FORM-ACTIVITY-MOVEMENT INTERACTION MODEL

Study of the Interactions between Urban Form, Allocation of Activities and Pedestrian Movement in Weimar, Germany

Dissertation

zur Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

an der Fakultät Architektur und Urbanistik
der
Bauhaus-Universität Weimar

vorgelegt von

Ing. arch. Martin Bielik
geb. am 02.12.1986

Weimar, 2020

Gutachter:

Vertr.-Prof. Dr. Sven Schneider

Assist. Prof. Dr. Pirouz Nourian

Tag der Disputation: 22.03.2021

Abstract

This dissertation investigates the interactions between urban form, allocation of activities, and pedestrian movement in the context of urban planning. The ability to assess the long-term impact of urban planning decisions on what people do and how they get there is of central importance, with various disciplines addressing this topic. This study focuses on approaches proposed by urban morphologists, urban economists, and transportation planners, each aiming the attention at a different part of the form-activity-movement interaction. Even though there is no doubt about the advantages of these highly focused approaches, it remains unclear what is the cost of ignoring the effect of some interactions while considering others. The general aim of this dissertation is to empirically test the validity of the individual models and quantify the impact of this isolationist approach on their precision and bias.

For this purpose, we propose a joined form-activity-movement interaction model and conduct an empirical study in Weimar, Germany. We estimate how the urban form and activities affect movement as well as how movement and urban form affect activities. By estimating these effects in isolation and simultaneously, we assess the bias of the individual models.

On the one hand, the empirical study results confirm the significance of all interactions suggested by the individual models. On the other hand, we were able to show that when these interactions are estimated in isolation, the resulting predictions are biased. To conclude, we do not question the knowledge brought by transportation planners, urban morphologists, and urban economists. However, we argue that it might be of little use on its own.

We see the relevance of this study as being twofold. On the one hand, we proposed a novel methodological framework for the simultaneous estimation of the form-activity-movement interactions. On the other hand, we provide empirical evidence about the strengths and limitations of current approaches.

Table of Contents

1 Introduction

When in 2009, the number of people living in cities surpassed the number of people living in rural areas, humanity became an urban species. Even though there is no single definition of what urbanity means, as the global population grew by 7.1 billion in the past 300 years, we can say that the world became a denser place. Not only do we live closer to each other than ever before, but cities also changed the activities we go about. In 1700, with 90% of the rural population, 60% of French citizens devoted their lives to single activity – agriculture. In 2012, with 80% population living in cities, the fraction of the workforce employed in agriculture dropped to less than 3% (World Bank Group, 2018). The effect of urbanization is that as everything gets closer, it becomes easier to move toward a larger variety of activities, which in turn boost interaction, communication, and innovation. In other words, “Cities were created to bring things together” (Speck, 2012, p105) with urban form as the dominant force affecting where we go and what we do.

Through the glasses of human history, urbanization looks like a force of gravity, pulling more and more people closer together as time goes on. And like the forces of nature, urbanization is neither good nor evil. As urban centers of the 18th century unleashed the scientific revolution and renaissance in philosophy, the same places became a scene of unprecedented human suffering (Roberson, 2016). When Ruskin mentioned “the great cry that rises from all our manufacturing cities” (Ruskin, 1853, p165), he was not talking in metaphors but offering an accurate depiction of despair brought upon the urban dwellers of the “flourishing” industrial cities.

As urbanity turned out to be a double-edged sword, the urban planners were commissioned to skillfully master it. Their task is to devise plans which unleash the positive and contain the negative aspects of bringing ever more people into limited space. "A city is not an accident but the result of coherent visions and aims" (Krier, 2009, p101), with urban planners being the driving force behind these visions and their realization. When Henri Ford introduced in 1908 the Ford Model T as the first affordable automobile, urban planners saw it as a chance to turn the “passel of people, packed in a pot like pickles” (Detzer, 2002) into “Space and light and order” (Le Corbusier, 1923). A grand vision which only a few decades later has been termed as “grand fiasco” (Blake, 1977).

The difficulty of modern cities was the notorious inability of urban planning to deliver on its promises. When instead of the “Space and light and order,” the results were “sprawl and smog and social unrest,” the impression arose that each new plan brings more problems than it was supposed to solve. The deadly efficiency of modernist urban planners in restraining human activities and movement is best summarized in the title of Jane Jacobs 1961 seminal book “The Death and Life of Great American Cities.”.

In essence, the attempt to eradicate the negative externalities of urbanity often crippled the reason why people originally moved to the cities. Or, as William H. Whyte puts it, "it is difficult to design a place that will not attract people. It is remarkable is how often this has been accomplished" (Whyte, 1988, p109). The ease of movement and ability to reach a wide range of activities were replaced by congestion and monofunctional neighborhoods. This is not to say that all cities have the same problem, and all urban plans are disastrous. However, due to the long-term effect of urban planning on billions of humans and their environment, the blatant failure to understand and predict the impact of planning decisions is hard to accept.

When searching for the causes of these failures, Jacobs identifies the lack of scientific rigor in urban planning as well as its tendency to religiously adhere to grand visions rather than to empirical evidence as the main source of trouble. "As in the pseudoscience of bloodletting, just so in the pseudoscience of city rebuilding and planning, years of learning and a plethora of subtle and complicated dogma have arisen on a foundation of nonsense" (Jacobs, 1961, p13).

The difficulty of urban planning is that there is no such thing as a coherent science of urban planning. The complexity of planning a city requires the joined effort of many professionals coming from different fields of expertise, using different methods and tools. Consequently, they often focus on different goals and propose different solutions, similarly as "the planning of the automobile city focuses on saving time while planning for the accessible city focuses on time well spent." (Cervero & Center, 1988, p2). As a result, even half a century after Jacobs wrote her harsh critique, and urban planning remains condemned as "anarchic, insubstantial, and full of internal inconsistencies" (Cuthbert, 2007, p177).

1.1 Research Gap and Problem Statement

The eclectic nature of urban planning combining different disciplines and individuals without clear methodological framing often results in a situation where the answer we get depends on the experts we ask. When it comes to understanding decisions on urban form and their effect on how people move and where activities take place, we find urban economists, urban morphologists, and transportation planners each proposing their own explanation.

The configurational school of urban morphology known as Space Syntax considers movement to be a direct product of urban form and its configuration. Its bold claims that movement and consequently the allocation of activities can be explained by looking at urban form only have been repeatedly criticized for lack of methodological rigor (Sevtsuk, 2010) and conceptual inconsistencies (Ratti, 2004). On the other hand, the professionals concerned with the topic of transportation planning tend to consider human movement as a product

of the allocation of activities. The urban form plays only the role of an interface connecting the origins and destinations of movement, and on its own, it does not explain anything (Barceló, 2010).

Finally, when urban economists tend to explain how activities allocate across space, they usually pay little attention to urban form or movement. The driving force behind their spatial distribution is considered to be economic interaction (Fujita et al., 1999). In other words, to explain why activities allocate at a given address, one must look at the neighboring activities.

As different urban planning professionals propose distinct models each focusing on specific parts of the interaction between the urban form, activities, and movement while ignoring others, we see the same variables being considered by some experts as dependent, by others as independent or being dismissed altogether. Even though there is no doubt about the advantages of simplicity and focus of each of the approaches, it remains unclear how they combine and what is the cost of this fragmented approach in terms of their reliability and validity.

We argue that cities are not built out of independent elements but are instead composed of a complex interconnected network of relationships where “something happens because something happens because something happens” (Gehl, 2014, p50). Under those circumstances, we expect a model that simultaneously considers all interactions between urban form, movement, and activities to bring essential advantages over each of the three disciplines' isolated approach. The essential question is, if one can meaningfully understand the interactions between urban form, movement, and allocation of activities without considering them simultaneously.

1.2 Aim of the Study

The overall topic of this study revolves around our understanding of the interactions between urban form, activities, and movement. In the context of different approaches focusing on specific aspects of the form-activity-movement interaction while ignoring others, our goal is to formulate and test a joined model considering all interactions simultaneously.

The aim of the joined interaction model is twofold. On the one hand, we expect the joined form-activity-movement interaction model to provide insights reaching beyond the narrow scope of the individual models. On the other hand, the richer, more comprehensive joined model is assumed to outstrip the accuracy of the individual models. Moreover, by comparing the outcome of the joined and the individual models we are not only able to assess their trade-off between accuracy and simplicity but, more importantly, test their bias and validity. By doing this, we seek to give the practitioners and decision-makers clear recommendations about how and when to use simpler individual models.

1.3 Scope of the Study

We explicitly limit the scope of this study to the investigation of the long-term, high-impact planning decisions on urban form and their effect on movement and allocation of activities. We argue that our ability to estimate their impact is of particular importance as they determine potentials and qualities which are only hard to change once build (Al-Sayed and Penn, 2016). Therefore, we restrict the broad notion of urban form to buildings and street network configuration as the most stable, long-lasting artifact of urban planning (Marshall, 2005).

When it comes to movement, we put a strong emphasis on one mode of transport - walking. Since cars shaped the 20th-century city, they are also blamed for its flaws and problems. On the contrary, it became clear that walking is an almost universal answer to urban problems ranging from individuals' health and social relationships to global warming (King & Murphy, 2017). The walkable neighborhood goes hand in hand with higher economic value, an increase in life quality, a decrease in crime rates, and a reduction of carbon dioxide emission. As more people seek to live in such environments, we see a growing number of walkability assessment tools (e.g., WalkScore, Walkshed) informing the location choice of tenants as well as the price demanded by the landlords. As a result, our understanding of the benefits of walkable environments and our ability to measure it improved in recent years. However, it remains unclear how to create such an environment. In specific, we have only limited knowledge about how the long-term decisions on the urban form (i.e., the configuration of street network and buildings) affect pedestrian behavior. We argue that after a century of research on motorized traffic, it is time to shed more light on how walking is affected by the urban form and allocation of activities.

Finally, the notion of activity is based on the concept by which the purpose of the movement is not to reach a destination but rather to fulfill the individual needs (e.g., hunger) by performing particular activity (e.g., visiting a restaurant). Even though it is closely related to the concept of land use, the difference between both is well expressed by Mitchell Silver's rhetorical question, asking, "who visits downtown to see land uses?" (CUN, 2017). The simple answer is "no one," and thus, the concept of land use is of little help when the goal is to explain the movement. Consequently, in this study, we focus on what people actually do at any given location rather than on how planners imagined the location to be used.

We must note that all interactions between urban form and human behavior are studied on the aggregated level. This simply means that instead of explaining the choice of activities and how to reach them for specific individuals, we try to understand these effects on the level of the whole population. We recognize the impact of demographics and individual preferences on travel behavior and the allocation of activities. However, we argue that understanding the aggregated effect of urban form on human behavior is better aligned with

the reality of urban planning. Not only are the individual urban dwellers and their preferences usually unknown, the longstanding character of urban form demands the ability to accommodate various needs as people come and go.

1.4 Review of Chapters

This study is presented in six chapters, accompanied by 19 appendices. After the introduction, we continue by reviewing the literature on representations and models of urban form, pedestrian movement, and allocation of activities. The individual topics are examined from multiple perspectives discussing the approach of urban economy, urban morphology, and transportation planning. In addition, we briefly review the general concepts of model building as the central topic of this study. We conclude the chapter with an overview of all current approaches and synthesize them into a joined form-activity-movement interaction model.

As next, we formulate two main research questions and a series of related hypotheses. We address the validity and performance of the models estimating the effect of urban form on movement in isolation by comparing them to the joined model.

In the fourth chapter, we present the design of the empirical study conducted in the mid-size German city of Weimar. We discuss the empirical data and methods devised to test the research hypothesis. To calibrate the movement simulation and activity allocation model, we conduct a series of empirical studies and summarize their outcome.

In chapter five, we present the results of hypothesis tests addressing the validity of the models estimating the effect of urban form on movement in isolation by comparing them to the joined mode.

Finally, we summarize the empirical study and discuss the results of hypothesis testing and in terms of its relevance to the field of urban planning. In specific, we conclude when different models can be considered useful and what are the risks and benefits of the individual and joined approach.

The main part is accompanied by 19 appendixes. They elaborate on the topics covered in the main part by offering greater detail and additional materials. By outsourcing a significant portion of the content into the appendix, we aim to improve the readability of the main part while maintaining the reproducibility of the study.

2 Literature Review

2.1 Model Building

Urban settlements are a complex product of many factors and simultaneous interactions between countless agents, making them notoriously difficult to understand. In natural and social sciences, one particularly successful approach to the study of such complex phenomena is model building. This research explores the interaction between movement, activities, and urban form through formal mathematical models. We start by defining the term "model," how it is built, and, most importantly, how it is tested.

Unlike the physical notion of the model used in the applied field of architecture or urban planning, the formal mathematical model used in this research context is an abstract concept capturing our knowledge about the world. At its most general level, the formal model is just any representation of something (Mitchell, 1993). Following the definition of Minsky (Minsky, 1968, p1), we use the term model in the following sense: "To an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A." In this sense, a model is useful tool which let us "answer a question about a hypothetical experiment without actually performing it" (Minsky, 1968, p1).

It is important to realize that any model is based on simplification and requires assumptions about the represented phenomena. Otherwise, we would not talk about the model but about a copy. Based on who is creating the model and what she is trying to answer, different simplifications are taken, and different assumptions are made. Consequently, there is no universal model of a system or phenomena superior to all other models. In this chapter, we discuss the assumptions and beliefs shaping the models used by different scholars regarding the study of pedestrian movement and the allocation of activities. As we can say that there is no single best model, there is also no single true model. As George Box famously put it, "all models are wrong, but some are useful" (Box, 1979, p2). In this study, we focus specifically on quantifying how wrong different models are and under which conditions we can consider them useful. Finally, based on existing models, we propose an alternative model combining the strength of different perspectives and mitigate their weaknesses.

According to Daellenbach (1995), models consist of three major components: a) Structure, b) Process, and c) relationships between structure and process. Dellenbach defines the structure as those elements of the model, which are static and do not change in the predefined time frame. In our context, we treat the elements of urban form such as buildings and streets as structure. The processes are defined as activities within the structure that undergo a change. We consider movement or allocation of activities as dynamic processes

restrained by the frame given through the structure of the urban form. The part which is notoriously difficult is defining the relationships between the structure and the process and between the individual processes.

As shown in the following literature review, the choice of elements of structure and processes and the preexisting knowledge about their relationship depends on the training, experience, and available resources. As a result, "there is always a degree of arbitrariness present in the model-building process" (Barceló, 2010, p2). Under those circumstances, regardless of the choice of the model components, the task of the analyst is to formulate and test the hypotheses on how the system works (modeling hypotheses).

When it comes to validation and assessment of the formal mathematical models, the question of how useful is a given model can be then evaluated in terms of its following properties (Figure 1):

a) Accuracy/Error – measuring how wrong is the model in a specific case
b) Bias – reflecting the average error of the model
c) Consistency – capturing if the model gets better as we calibrate it with more data
d) Efficiency – measuring the variation of the error

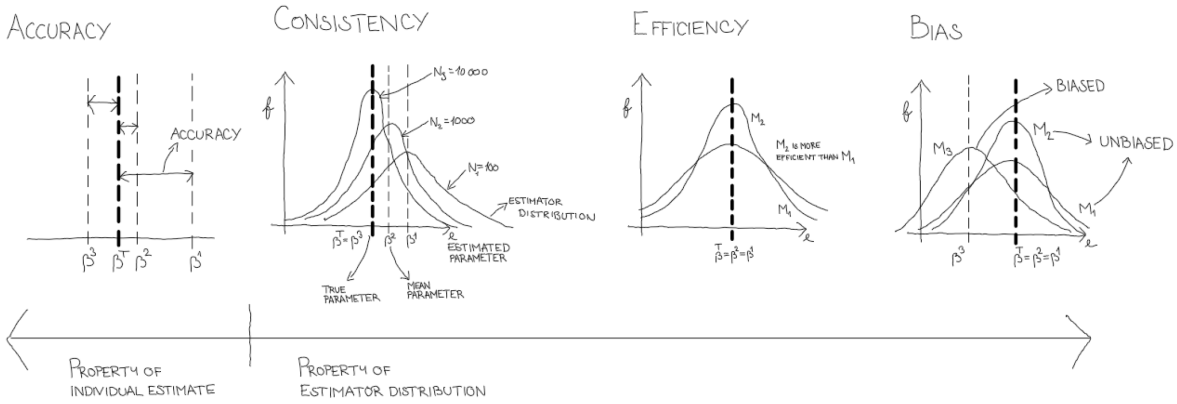


Figure 1. Regression model properties: Accuracy, consistency, efficiency, and bias of estimator (source: Author)

All these model characteristics are routinely used to evaluate a specific type of formal mathematical models – regression models and their estimators. Each represents a different property of the model performance and has to be considered when evaluating its usefulness (Amemiya, 1985).

Accuracy

It is important to realize that all models will have some degree of error. The reason is the inherent incompleteness of any model, as already discussed before.

Bias

Bias is the property of a model based on the distribution of the error over many experiments. Models that are correctly specified might be in each individual case wrong, but on average, they are correct. We call such models as unbiased models. On the contrary, if the expected outcome of the model (i.e., its mean estimate) is wrong, such a model is biased. The bias of a model suggests that not only part of the reality was ignored. More importantly, it means that the part which has been included in the model is not represented correctly. In general, we accept the error and try to avoid the bias (Lehmann, 1951).

Consistency

Consistency is the asymptotic property of the model. It measures how the amount of data used to calibrate the model influences its error. An asymptotically consistent model is perfectly on target as the number of data points increases infinitely (Sober, 1988). On the contrary, the model is inconsistent if it does not improve as the data sample increase.

Efficiency

Finally, if two models have different efficiency, their error varies differently over many experiments (Everitt & Skrondal, 2010). One might always be similarly accurate, while the other might sometimes be entirely on target and another time completely wrong.

2.1.1 Summary

The role of the model is to formalize what we know about the world. It allows us to formulate and test hypotheses and conduct experiments. The assessment of the model revolves around multiple dimensions, with the ideal model being unbiased, consistent, very efficient, and with error always close to zero. However, for a model to be useful, it does not have to be ideal. Based on the research question, we might accept a various amount of error or efficiency. Similarly, based on the amount of available data, we might not even care about consistency. Finally, there is one characteristic of the model we always care about - the bias, as this might cause the model not only to be imperfect but also pointing in the wrong direction.

2.2 Urban Form

This chapter discusses and defines the term urban form as a central element of this study. We start with a review of the core concepts of discipline devoted to the study of urban form – urban morphology.

Urban morphology is concerned with "Classifying, describing and quantifying the urban form" (Pont & Haupt, 2007). The primary source of knowledge in urban morphology is the recognition of patterns (Scheer, 2015). These patterns are what urban morphologist also call urban tissue. Karl Kropf, in his paper 'Urban tissue and the character of towns' (Kropf,

1996) defines urban tissue as an organic whole observable at different levels of resolution. The higher the level of resolution, the greater the detail of what is shown and the greater the specificity of morphological description (Kropf, 1996). To deal with the complexity of urban tissue, urban morphology adopted a hierarchical view of the city, structured according to a set of three fundamental physical elements: streets, plots and buildings (Oliveira, 2016).

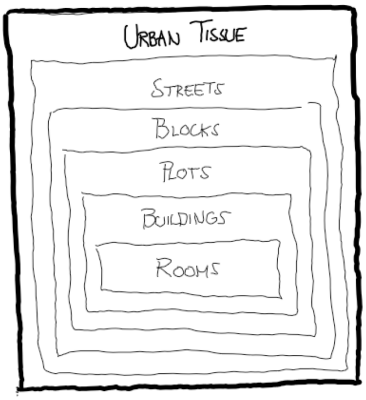


Figure 2. Hierarchy of fundamental elements of urban form (adapted from Kropf, 2016)

2.2.1 Streets

Streets are at the top of the hierarchy of the elements of the urban form. They define the space occupied by all other elements of urban form and are connecting everything with everything else (Marshall, 2005). Streets are the most stable element of urban form that defines and connects the urban tissue. "While the physical process of city building is something that 'takes time' involving permanent transformation—it has a past, a present and a future—the streets system of a city is the one that offers greater resistance to this process of urban transformation, attaining a great temporal stability" (Oliveira, 2016, p15). One of the most evident demonstrations of the longevity of street patterns was their ability to survive even the complete destruction of many urban centers after the air raids of the second world war. "The reconstruction work often followed existing streets, a phenomenon that included city planners to comment that we are building around own sewers" (Hofmeister, 2004, p4).

2.2.2 Plots

The plots are the abstract system of land subdivision defined by the street block. They are separating the private and public domain and delineate the structure of land ownership. When compared to streets and buildings, the plot system does not require any materialization to exist. It becomes visible indirectly, through other elements of urban form, most notably the buildings. Despite the abstract nature of the plot system, it is one of the essential elements of urban form with considerable stability over time. Urban morphologists found that the subdivision and plot amalgamation are rather rare, which means that the

choices made in the early stages of the urbanization process have a long-lasting effect on the urban forms built in the city (Conzen, 1988). Oliveira concludes that "although the city suffers many kinds of disturbances over its life—such as wars, fires, earthquakes, tsunamis, to name just a few—that could be used as a pretext to erase the pre-existing plots system (or parts of the plots system) and to create a new plot structure, the truth is that, in most of the cases, this does not happen and the pre-existing plots system is maintained" (Oliveira, 2016, p24).

2.2.3 Buildings

Buildings are the most visible element of urban form, central to urban life as they host a large portion of human activities. The shape and distribution of buildings are determined by the system of plots and streets and has far lower stability over time. In general, urban morphologist distinguish between two types of buildings. Based on the building form and building utilization, they differentiate between the ordinary and exceptional buildings. The ordinary buildings constitute the majority of the total building stock in the city and host most human activities such as accommodation, work, commerce, and services. The similarities between buildings within this type are more substantial than their differences (Oliveira, 2016).

On the other hand, the small set of exceptional buildings display a high level of variance in the form with utilization limited to a few specific activities. Each building of this type is easily distinguishable from the rest of the building stock of the city. In general, urban morphologists trying to understand the how and why of urban form focus predominantly on the ordinary-buildings rather than a small minority of exceptional buildings (Pinho & Oliveira, 2009). As Renn puts it, "The mark of a great city isn't how it treats its special places – everybody does that right – but how it treats its ordinary ones" (Renn, 2013).

2.2.4 Approaches to Urban Morphology

Even though the urban morphologists agree about what they study, there is considerable debate within the field over how urban form should be studied (Gauthier & Gilliland, 2005). As a result, all urban morphologists collect data, however, not necessarily the same type of data. Moreover, when they analyze and synthesize the data, they are not necessarily using the same methods and coming to the same conclusions.

Elements and Resolution

Different scholars focus on different elements of the urban form and study them at a different resolution. For example, a street can be represented as a simple line characterized only by its length and connection to other streets, or we can add the second and the third dimension resulting in a much more nuanced set of characteristics. The resolution of the study is mostly an inverse function of its scale. As the scale of the study increases (e.g., from neighborhood

to city or region), the resolution of the elements of urban form decreases (e.g., from building forms capturing openings and roof structure to simple extruded volumes).

The exact definition of how to represent an individual element of urban form at each level of resolution is an ongoing subject of discussion in the field of urban morphology. Nevertheless, In recent years considerable effort was invested into defining formal standards that make the use and exchange of urban data between morphologists, geographers, planners, developers, and policymakers more accessible. The resulting international CityGML standard for representation and exchange of 3D city models issued by the Open Geospatial Consortium (OCG) defines five discrete resolution levels termed Level of Detail (LoD) (Kolbe, 2009). Each LoD consists of different elements of urban form and defines how these should be represented (Figure 3). From a conceptual standpoint, each LoD reflects specific application requirements and allows the same urban form element to be simultaneously represented differently across the LoDs. In the following, we briefly define the main features of each LoD, as summarized by Gröger & Plümer (2012).

- The LoD0 captures the urban form at the regional level and represents urban form at the largest scale and the smallest resolution. Essentially it is 2.5D representation of urban form described by polygons embedded in 3D space. Buildings and plots are represented by a single horizontal polygon, and transportation objects are generalized by linear structures in LoD0.
- The LoD1 target citywide applications and represent volume objects (e.g., buildings, vegetation) in a generalized way as extruded blocks vertical walls and horizontal 'roofs.' Transportation features are represented as 2.5D surfaces.
- The LoD2 zooms in to the city district level. Buildings are represented as extruded volumes with prototypic roof shape, thematic ground, wall and roof surfaces, and structures such as balconies and dormers. Natural elements and vegetation objects also are represented with simple 3d geometric shapes. The representation of transportation objects is enhanced by the explicit representation of traffic areas by their functional purpose.
- The LoD3 captures urban form at the level of architecture with the most detailed level for the outermost shape of objects. Buildings, as well as transportation and land use objects, are represented by their outer boundaries in a very detailed way, including openings and auxiliary construction.
- Finally, in LoD4, interior structures (rooms, etc.) are added, resulting in a complete architectural model with a detailed representation of its interior and exterior form.

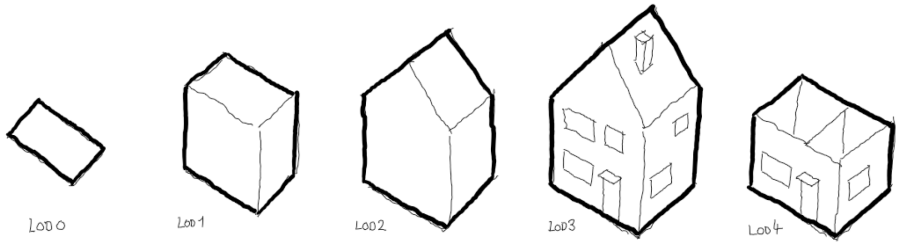


Figure 3. Five Levels-of-Detail (LoD) provided by CityGML (Adapted from: KIT Karlsruhe, K.-H. Häfele, Gröger et al., 2012).

Methodological Approach

Besides the fact that different morphologists focus on different urban form elements, a more fundamental divide can be drawn according to the approach or method used to study these forms. Gauthier & Gilliland (2005) identified two criteria for the classification of the aims and means of different scholars. On the one hand, they distinguish between cognitive and normative approaches. On the other hand, between what they call externalist and internalist contributions.

The cognitive and normative approach characterizes the primary heuristic purpose of the study. In other words, what the author wants to convey. On one side of the spectrum is the cognitive approach following the intention to analyze and explain existing morphological patterns. In contrast, on the other side, we find the normative approach aiming to prescribe and define how future urban form should be planned.

The second axes along which Gauthier and Gilliland categorize the approaches to urban morphology is distinguishing between internalist and externalist view on how urban form came to being. Internalists consider urban form as a more or less independent system that can be explained on its own. In other words, they argue that we can understand the patterns of urban form at any point in time just by looking at the urban form at previous points in time. On the contrary, for the externalists, the urban form is a "dependent variable or passive product of various external determinants" (Gauthier & Gilliland, 2005). From the externalist perspective, the urban form is a product of political, economic, and social forces.

Traditionally, the different schools of urban morphology are categorized by their historical and geographical roots. Oliveira, Kropf, and Sheer identify three central approaches to urban morphology: Process typological (Italian), Historic-geographical (British), and Configurational (Space Syntax). However, when looking at the data, resolution, and methodology of the individual scholars and their studies Gauthier and Giilliland (2005) find out that the variance inside of the group is often higher than the variance between the groups. They mapped the seminal works in urban morphology on a two-dimensional plane dividing the individual approaches into four categories: a) Cognitive & Internalist, b) Cognitive & externalist, c) Normative & Internalist, and d) Normative & Externalist (Figure

4). This categorization reveals a wide range of attitudes with a modest preference towards the Cognitive & internalist approach. It is often the case that not only the protagonist of the same school use different approaches (e.g., Conzen and Whitehand), but also the same scholar is adapting her methodology based on the research question being answered (e.g., Lynch, Rapaport).

Figure 4. Mapping approaches in urban morphology (Adapted form Gauthier & Gilliland, 2005).

2.2.5 Summary

Drawing upon the common characteristics and specificities of individual approaches to urban morphology, it becomes clear that it offers guiding principles but no universal approach to defining how the urban form should be represented and studied. The approach has to be adapted to the research question, scale, resolution, and the overall aim of the study.

The aim of this study to investigate the effect of urban form on pedestrian movement and allocation of activities implies the study of a city as a whole. The result of the citywide scale implies the adoption of CityGML LoD1 with the street system being represented as a line network, plots as polygons, and building as simple volumetric extrusions.

From a methodological perspective, the clear aim of this study is to expand our knowledge of urban form, and thus it clearly leans towards the cognitive part of the spectrum. We argue that we must understand how cities are in the first place before prescribing how they should be. As a result, we discuss how our findings can inform urban planners and policymakers. However, we do not elaborate on how the resulting urban form should look.

Regarding the internalist-externalist divide, it must be noted that the aim of this study is not to explain how the urban form came into being but rather what are its characteristics and how these influence the movement 'to' and the allocation 'of' human activities. This renders the externalist approach less favorable as it aims the exact opposite - to explain how human behavior affects the urban form. On the contrary, to most distinct of the cognitive internalist – the configurational school around Bill Hillier, also known under the term "Space Syntax," with the main focus on urban form and human movement, makes for an ideal methodological cornerstone for this study.

2.3 Pedestrian Movement

Understanding how people move is a key aspect of urban planning and policymaking. In this research, we limit the scope of our investigation to pedestrian movement as "the silver bullet to health and wellbeing" (King & Murphy, 2017). In recent years, the research on walkability contributed significantly to our understanding of why people walk. Multiple walkability measures have been developed, combining the individual needs of pedestrians and what the environment has to offer in terms of a) where to walk and b) how to get there. Alfonzo (2005) describes a hierarchy of pedestrian needs with lower needs, which must be satisfied before satisfying the higher-level needs (Figure 5). This concept, closely leaning on the famous Maslow's pyramid of needs (Maslow, 1943), starts with the feasibility as the basic need. It refers to the individual impairments of a pedestrian (e.g., age or disability) that influence the decision whether or not to walk. If an individual is physically able to walk, the next most important factor identified is the accessibility of walking destinations. Only then the higher-order needs such as safety, comfort, and pleasurability come into play.

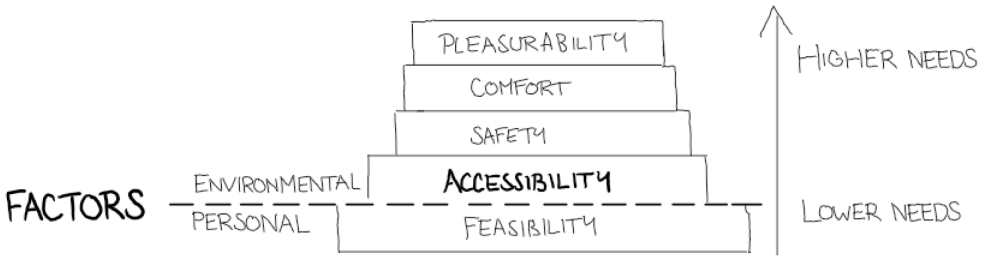


Figure 5. Hierarchy of pedestrian needs (Adapted from Alfonzo, 2005).

When measuring walkability, different researchers focus on different environmental characteristics at a different level of resolution. This can be attributed to the wide range of applications of the walkability measures aiming at explaining a physical activity, health, environmental footprint, real estate values, crime, or social cohesion. Just to mention a few, the Walkscore (WS) and the Walkability Index (WAI) are arguably the most prominent indices of walkability (Reyer et al., 2014). WS is based on fully automated online walkability assessment and became recently one of the most widely applied environmental quality assessment methods accounting for more than 20 million evaluations per day (walkscore.com). The walkability of any address is assessed via the accessibility of several categories of nearby amenities. Points on the scale from 0 to 100 are awarded based on the distance to amenities in each category. Amenities within a 5- minute walk are given maximum points. A decay function is used to give points to more distant amenities, with no points given after a 30-minute walk. WAI takes a similar approach focusing mostly on diversity and the accessibility of walking attractors offering freely available GIS implementation of the measure (Frank et al., 2010). The main advantage of WS is that most of the data required for the assessment is readily available online. Thus, the assessment can be fully automated.

Other measures, such as Walc Institute's Walking Audit Survey Tool or the State of Place index, take a more detailed look at walkability, accounting for more than 200 features. While it is assumed that the more variables are included, the higher the precision of the assessment, accounting for higher-order needs such as comfort and plasurability comes at a high cost of laborious data acquisition and difficult evaluation. In contrast to the distribution of land use and amenities, the data on the microscale characteristics of the built environment (e.g., sidewalk width, rowdiness, crossing type) are being collected predominantly manually since no standardized database exists.

Additionally, even after the data on multiple variables have been collected, it remains unclear how to aggregate it into a single walkability measure. The general aggregation approach shared by most walkability assessment methods is the simple sum of the normalized and weighted individual factors resulting in the final score (Lefebvre-Ropars et al., 2017). Where the methods differ is how the weightings are defined. While many measures

exist today (more than 80, according to Vale et al., 2015), few of the indices proposed in the literature have been validated against actual survey data. The weighting of these few are often based on expert knowledge (Ewing & Handy, 2009) and do not account for the collinearity and interactions between the factors. By treating them as independent and ignoring the complex urban reality (Frank et al., 2010), the contribution of the large multivariate walkability measures remains questionable. As Grasser and colleagues (2013) pointed out in their review of walkability measures, there is a shared understanding about which walkability factors are important. However, their weightings are highly inconsistent.

We have to mention that walkability research, in general, does not aim at modeling movement. The measures of walkability are merely informing movement models about the pedestrian needs and environmental characteristics relevant to one's decision whether to walk or not to walk. They usually do not consider the movement as a complex Spatio-temporal process consisting of origins destinations and routes connecting them and are limited to an assessment of the existing environment with prior knowledge of the distribution of activities (e.g., walking attractors) and detailed characteristics of the path leading to them.

Movement models in general, and pedestrian movement models in specific, illustrate the much broader statement that models differ in their approaches depending on the modeler's purpose and resources (Minsky, 1968). If the goal is to explain movement from an aggregated point of view, we talk about macroscopic models, whereby if the aim is to describe the movement of individuals, we talk about disaggregated – microscopic models (Helbing, 1998). The focus of this research lies in the understanding of the interaction between the urban form, movement, and activities. We model these interactions on the scale of the system rather than the individual. Thus, in the following section, we focus on the review of the aggregated movement models.

Based on the assumptions about who, how, and why is moving, we talk about two major groups of models looking at movement either from the a) transportation planning perspective (TP) or b) the perspective of configurational urban morphology (CUM). In the TP, the focus lies on the process and the agents of the movement. In the CUM, the elements of the urban form that facilitate the movement come into the foreground. Each approach requires different information and is based on different assumptions. We discuss how each model in terms of its advantages, limitations, and usefulness for the purpose of this study.

2.3.1 Transportation Planning Approach

The prevailing approach to travel in transportation planning literature is the human activity approach (Fox, 1995; Jones, 1983). It is based on the concept by which the purpose of individual travel behavior is not to reach a given location but rather to fulfill the individual needs (e.g., hunger) by performing particular activities (e.g., visiting a restaurant). Since

most human activities are temporally and spatially constrained, travel is seen as a mechanism for overcoming these constraints (Hägerstrand, 1970). Empirical studies of human activities and their Spatio-temporal constraints suggest that some activities are more fixed or bound than others (Cullen & Godson, 1975; Doherty, 2006). Gehl (1987) extends this concept by differentiating activities by the degree of personal freedom of choice. He proposes four categories of activities starting with the most constrained a) necessary activities and continuing to b) contracted, c) committed, and d) free activities. The major contribution of this conceptual framework is the realization that the same environmental change might have a different effect on travel behavior based on the type of activity being accessed. If the access to necessary activity is decreased (i.e., it is further away), the effect is increased travel time and traffic volume. On the contrary, in the case of free activity, the same change might result in a decision not to travel and thus decrease the travel time and traffic volume (Elldér, 2014).

Given all these points, the movement models in transportation planning consist out of three components a) travel network, b) travel demand, and c) travel route (Barceló, 2010). In other words, how people move (i.e., travel route choice model) is seen as a product of the interaction between the personal preferences and needs (i.e., activity demand model) and distribution and configuration of urban form and its characteristics (i.e., the travel network model) (Figure 6). Consequently, the model of the overall movement in the system is the result of interaction and aggregation of the individual agents (e.g., person, household, block, neighborhood) and their movement routes.

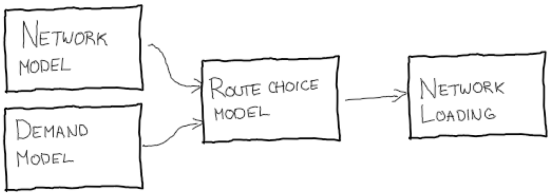


Figure 6. Conceptual scheme for mobility based movement models (Adapted from Barceló, 2010).

The major challenge in the model building process is the collection of environmental and behavioral data. While the geoinformation technology and the internet made the collection and sharing of large data sets about the urban environment ever more accessible, the data on personal needs and resulting travel preferences remain notoriously difficult to collect, or it simply does not exist yet (modeling future urban systems).

Since travel needs are often unknown, the common practice is to look at travel options instead. In such a case, we do not talk about mobility - the realized travel, but about accessibility – the potential travel (Curtis & Scheurer, 2010). In simple words, it is assumed that the more travel destinations can be accessed, the more travel will be eventually realized (Figure 7). This assumption is largely based on the notion of travel as derived demand,

which means that the majority of trips are taken to reach a particular destination and are not considered as a purpose on their own. Even though the concept of travel as derived demand is largely true in the case of automobile travel, it might be less so in the case of walking (Handy, 2002). From the empirical travel data, we found that 98% of car trips in Germany can be classified as derived demand as opposed to 80% for pedestrian trips during the same time period (MiD, 2017).

Using accessibility as a proxy of mobility might be useful in some cases, but as Handy (2002) points out, the relationship between mobility and accessibility is more complicated than this. She points out that the policies to increase mobility will generally increase accessibility as well by making it easier to reach destinations. However, it is possible to have good accessibility with poor mobility. One can imagine a "congested neighborhood where residents live within a short distance of all desired destinations, which has poor mobility but good accessibility" (Handy, 2002). Additionally, based on the type of travel needs, the demand for travel gets eventually saturated, and the increase in travel potential above this threshold will not result in a further increase in realized travel.

In spite of these caveats, the accessibility approach to movement modeling is widely considered as a useful approximation of mobility with the clear advantage of simpler data collection as the information about the allocation of travel opportunities can often be easily acquired from publicly available data sources. For the case study area used in this research, the travel opportunities can be assessed from governmental (e.g., Geoproxy service maintained by the regional authorities) and non-governmental (e.g., OpenStreetMap) data sources, while the data on the travel is not available at current times and must be collected manually. Due to the limited resources, the collection of the travel demand data required for the mobility-based movement model is not feasible at current times, and thus, we focus in the course of this review on the accessibility movement modeling as the best alternative.

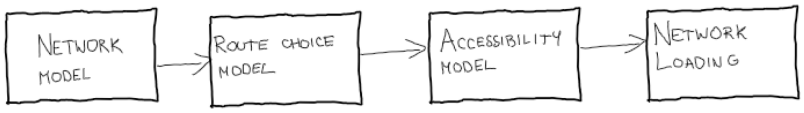


Figure 7. Conceptual scheme for accessibility-based movement models.

Network Model

The transportation system is composed of the origins, destinations (i.e., allocation of activities), and the infrastructure connecting them. These can be represented at different resolutions depending on the goal of the study. From the morphological point of view, the transportation infrastructure is represented as a street system, while the origins and destinations of movement are mostly represented as buildings. However, in some cases, the movement is not bounded to the streets, and it does not necessarily start or end in buildings. An example of the former is "the modernist plan replacing the street block by a new type

of urban landscape where buildings are completely disconnected from the street" (Oliveira, 2016). In the case of the origins and destinations of movement, approximately 20% of pedestrian movement in Germany is not driven by the need to reach buildings (MiD, 2017). In such cases, the movement is the purpose of its own. Nevertheless, for the context of European cities where this study is grounded, the empirical data on urban mobility suggest that a network model composed of the street system and buildings is a reasonable approximation of reality.

Route Choice Model

Estimating potential travel (i.e., accessibility) or realized travel (i.e., mobility) is all about quantifying relationships between locations. According to Tobler's first law of geography, "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p234). What it means is that essentially, when talking about spatial relationships, we are talking about distances. Measuring distance between any origin-destination pair is traditionally achieved by calculating the shortest path between them. These paths have been assumed to be "the result of minimizing procedures such as selecting the shortest path, the quickest path or the least costly path." (Golledge, 1995).

There has been an ongoing debate over what metric best represents the cost of travel. Different scholars advocate for different metrics for calculating the travel costs. In transportation planning, two major measures are dominating the debate over travel cost definition – the metric and the temporal distance (Porta, 2010; A Sevtsuk, 2010). Nevertheless, based on the individual, the mode of travel, and the geographical, environmental, or socio-cultural context, other factors such as slope, safety, air-pollution might play a significant role.

As a result, there is no definite universal set of rules for representing distance, and the decision is often based on the local expert knowledge and calibration done by the modeler.

Accessibility Model – Movement Potential

After identifying the routes connecting all possible origins and destinations of movement, the contribution of each of them to the total movement must be estimated. Even though everybody agrees that different destinations might attract a different amount of movement, there is much discussion going on about how to quantify these differences. In the following, we review the family of accessibility estimation methods used to measure the contribution of each destination to the overall movement potential.

The concept of accessibility was first introduced by Hansen (1959), who defined it as the potential of opportunities for interaction. As Batty concludes, "the accessibility is often seen as a measure of the cost of getting from one place to another, traded off against the benefits received once the place is reached" (Batty, 2009, p191). When measuring total accessibility

from one place to all others, we end up having an aggregate measure of how easy or difficult it is to realize all these opportunities from a given location.

Based on the point of view from which the accessibility is measured, Hanson (1995) and Miller (2007) differentiate between active and passive accessibility. The active accessibility represents "the easiness in carrying out activities (e.g., shopping, entertainment, education, work, etc.) for a subject located in a certain zone" (Cascetta et al., 2013, p118). On the contrary, passive accessibility captures the ease of being accessed. In this research, we measure accessibility to inform the movement model about the travel potential of individuals at a given location. Thus, we are looking at it from an active or personal point of view. This distinction is helpful as each of these two perspectives involves different methodological challenges, which will be discussed in the following section focusing solely on the active accessibility.

The definition of an active accessibility measure involves a number of interrelated issues such as the degree and type of disaggregation, the definition of origins and destinations, and the measurement of attractiveness and travel impedance (Handy & Niemeier, 1997). As a result, multiple alternative active accessibility measures exist, with no one being superior to the others. As Handy and Niemeier put it: "different situations and purposes demand different approaches" (Handy & Niemeier, 1997, p1181). They identify a trade-off between the complexity and precision and cost of calculation and the difficulty of interpretation. Therefore, the choice of an accessibility measure depends on the available resources and the aim of the analysis.

Before going into the specificities of the individual active accessibility measures, we would like to point out two general principles or requirements which they all must fulfill. Geurs and van Wee (2004) argue that regardless of the perspective of an accessibility measure, it should always:

a) relate to changes in travel opportunities, their quality and impediment: "If the service level (travel time, cost, effort) of any transport mode in an area increases (decreases), accessibility should increase (decrease) to any activity in that area, or from any point within that area." (Geurs & Wee, 2004, p130)
b) relate to changes in land use: "If the number of opportunities for an activity increases (decreases) anywhere, accessibility to that activity should increase (decrease) from any place." (Geurs & Wee, 2004, p130)

Based on these two principles, Handy and Niemeier (1997) have identified three distinct classes of accessibility measures commonly used in the transportation planning literature: cumulative, gravity, and utility-based accessibility. Each of the three measures can be characterized by a certain trade-off between complexity, accuracy, and ease of interpretation.

The cumulative accessibility, or also called isochronic accessibility measure, is the easiest to calculate but take the most arbitrary assumptions (Handy & Niemeier, 1997). It is defined as the sum of potential activities (destinations) available to an individual (origin) located in a given distance radius.

$$A_i = \sum_{j, i \neq j, d_{ij} < t}^{n} O_j \tag{1}$$

Here the O_j is the opportunity at location j and t is the distance radius threshold. The only information needed for this measure is the location of all destinations and the threshold defining the maximum acceptable travel distance (Bhat et al., 2002).

In spite of the ease of calculation and interpretation, the cumulative accessibility has been often criticized for "the lack of a behavioral dimension and the incapability to model the differences in the perception of near and far opportunities, i.e., opportunities are equal regardless of their cost and desirability for users" (Cascetta et al., 2016). The key element during the measure calibration process is the choice of the cut-off travel distance or time – the isochrone radius. This is a highly problematic and arbitrary decision as it considers the distance to all destinations inside of the isochrone as equal, despite the fact that actual travel times obviously vary among activities within the same isochronal line. Within the transportation planning, 400m or 5-minute walk is often used as a rule of thumb for the pedestrian movement radius (Vale & Pereira, 2016). However, the empirical studies show that the real walking distances are a) consistently greater than this and b) vary widely across different cities based on the type of public transport service, the reliability of the service, the opportunity being accessed, but also individual and socio-economical characteristics (Vale & Pereira, 2016). As a result, the definition of cumulative accessibility changes from study to study based on arbitrary decisions made by the modeler and make results difficult to compare.

The gravity-based accessibility measure extends the cumulative accessibility by addressing its core weakness – the fixed travel distance threshold. It is a product of the attractiveness of the destination O_j and the inverse of the travel distance d_{ij} required to reach it. It means that all destinations are considered, but those ones which are closer or more attractive are contributing more to the total accessibility index than those further away or less attractive (Hansen, 1959).

$$A_i = \sum_{j, i \neq j, d_{ij} < t}^{n} \frac{O_j}{f(d_{ij})} \tag{2}$$

The choice of the distance function and its parameters have a significant impact on the result of the accessibility measure (Kwan, 1998). In general, different contexts and modes of transport result in different functional forms between the distance and the willingness to travel (Figure 8). The negative exponential function has been the most widely used within transportation planning, indicating decreases in the exponent as trips become more important.

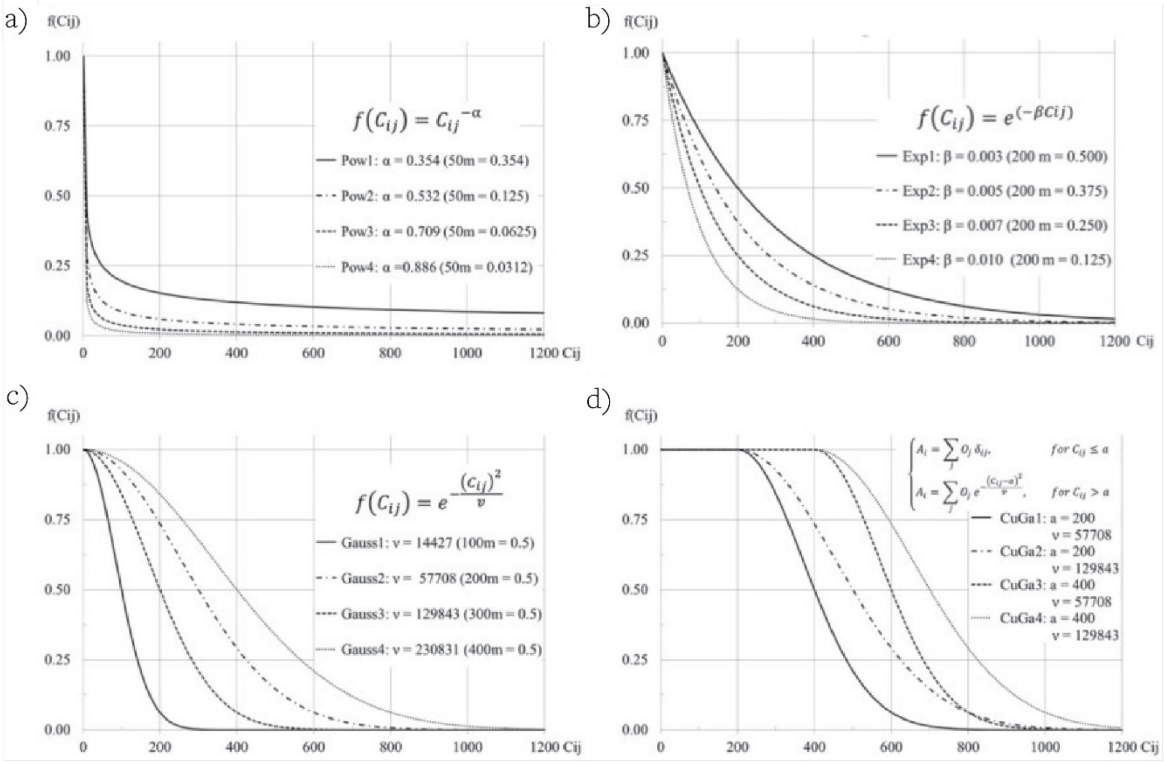


Figure 8. Impedance functions used in the pedestrian accessibility measures. (a) Power, (b) exponential, (c) Gaussian, and (d) cumulative– Gaussian (Source: Vale & Pereira, 2016).

Compared to the cumulative accessibility, the gravity-based accessibility is closely tied to travel behavior theory backed by empirical data on how humans move (Handy, 1992). In the process of calibration, the parameter value for the travel impedance function and attractiveness of destinations must be estimated. The common approach to the distance decay parameter calibration is the use of an empirical trip-distribution model (Bhat et al., 2002). The attractiveness is often represented as simply the amount of activity, measured by numbers of employees or square meters (Handy, 2009).

The utility-based accessibility as the last and most complex class of accessibility measures builds on the random utility theory. The core concept is based on a probabilistic approach to travel where the travel choice of an individual depends on the utility of that choice relative to the utility of all possible choices (Handy & Clifton, 2001). It is assumed that individuals "must choose one of a finite number of exclusive alternatives" (de Jong et

al., 2007, p878) by assigning utility to each of them and then searching the set and selecting the best possible option. Formalized by Ben-Akiva & Lerman (1979), its general form reads as the accessibility for an individual at location i being the expected value of the maximum of the utilities over all alternative spatial destinations j.

$$A_i = E\left[Max_{j,i\neq j,d_{ij}<t}\, U_j\right] = ln\left[\sum_{j,i\neq j,d_{ij}<t}^{n} \exp\left(V_j\right)\right] \quad (3)$$

These models are calibrated from travel survey data containing the information on each trip characteristics such as travel impedance, the attributes of both the destination and the traveler. These are included as explanatory variables in a utility function. The modeler can test alternative formulations of the utility function to find the one that best matches actual travel behavior. The advantage over the cumulative or gravity-based accessibility is that the relative importance of different factors is determined through the calibration and does not have to be arbitrary prespecified by the modeler. On the other hand, the price for it is the very demanding calibration and data collection process. The major criticism of this framework is the behavioral basis of its assumptions. It has been repeatedly questioned by psychologists and social scientists, pointing out that a) the activities found at different destinations are not necessarily mutually exclusive (de Jong et al., 2007) and b) that humans are notoriously bad at making rational utility-maximizing choices (Tversky & Kahneman, 1974). As a result, the accessibility measure based on utility might neither reflect the observed behavior nor the benefit of increased choices (Morris et al., 1979).

Network Loading

After estimating the travel potential for each individual agent or location in the network, the total network loading of the whole system can be modeled. In transportation planning, travel is conceptualized as a dynamic process with roads having limited capacity and individuals trying to minimize their travel time. As a result, the overall traffic loading model is more than a simple aggregation of individual trips and can be described as a dynamic traffic assignment (DTA) problem. It determines the time variations in link or path flows and is capable of describing how traffic flow patterns evolve in time and space in the network (Mahmassani, 2001).

Solving the DTA is based on the behavioral assumptions about how travelers choose their routes described by the dynamic user equilibrium principle. Ran and Boyce (1996) formulated it in the following terms: If for each origin-destination (OD) pair at each instant of time, the actual travel times experienced by travelers departing at the same time are equal and minimal, the dynamic traffic flow over the network is in a travel time-based dynamic user equilibrium (DUE) state.

Various algorithmic schemes implementing the DTA have been proposed using numerical, stochastic, and heuristic-based simulation model (Barceló, 2010). Without going into further details, the main take away is that the travel behavior of an individual is not considered in isolation but in interaction with other agents and their travel needs and the transportation infrastructure. Peeta & Ziliaskopoulos (2001) conclude that DTA movement modeling is a trade-off between accuracy on one side and the costs related to computational complexity and data acquisition on the other side.

2.3.2 Configurational Urban Morphology Approach (Space Syntax)

Space Syntax is often seen as a configurational school of thought of urban morphology (CUM) and can be described as a conceptual framework for understanding the social consequences of morphological configurations. Its origins can be traced back to the seminal work "Social logic of space" by Hillier and Hanson (1984) establishing the core pillars of Space Syntax: a) the human perception approach to the representation of urban form and b) the configurational approach to quantifying the properties of urban form (Lee & Ostwald, 2019). The former means that instead of representing the urban form through its individual elements such as streets and buildings, it is represented as humans perceive it through visual fields and lines of sight and movement. These urban form elements are then put into spatial relations, and their resulting configurational properties and their effect on human behavior are the main focus of the Space Syntax studies.

In his seminal paper, *Cities as movement economies* Hillier (1996) argues that one specific type of behavior –movement is the fundamental driving force of urban life. The social, economic, and environmental consequences of urban form are seen as a product of its configuration, which influences how we move around cities, which in turn influences everything else.

The Space Syntax method revolves around graph theory – a field of mathematics specialized in quantifying configurational characteristics of relational systems. Its origins date back to the seventeenth century and the work of swiss mathematician Leonard Euler first applying the concept of the graph when solving the puzzle known as the seven bridges of Königsberg. The problem was to devise a walk through the city that would a) visit all of its four landmasses divided by rivers and accessible only by bridges and b) that would cross each of those bridges once and only once. Euler's solution to the problem was based on the simplified mathematical model of relationships and possible movement routes composed of vertices (the landmasses) and the edges (the bridges) (Figure 9). This model, known as the graph, was rediscovered in the middle of the nineteenth century by geographers and later by urban morphologists due to its power to quantify and explain various social phenomena (Harary, 1960, Lee & Ostwald, 2019).

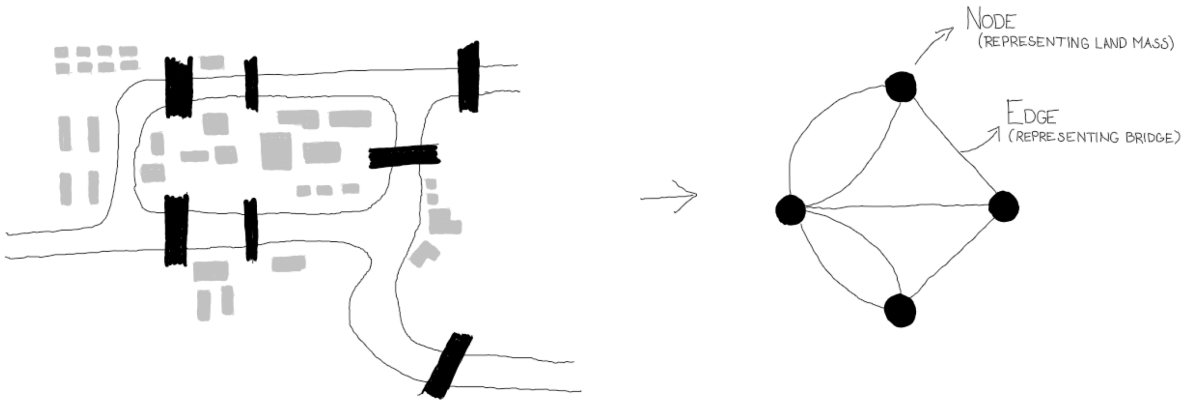


Figure 9. Euler's graph representation of the Königsberg puzzle.

As argued by Space Syntax scholars, their graph-based analytical model is able to explain 60 to 80% of the variance in movement flows as an effect solely of the street network configuration (Penn 2003). This model measures the centrality of each location (i.e., graph node) with respect to the rest of the system to estimate its potential to attract and guide movement. It must be mentioned that Space Syntax scholars treat their model as a model of urban form with the ability to explain movement instead of a model of movement based on urban form (Hillier, 1999). This conceptual standpoint is what makes the Space Syntax approach conceptually different from TP. However, when it comes to the integral elements of both approaches, we see a high degree of similarities between them. They are both based on a) network model defining the origins, destinations, and infrastructure of movement and b) route choice model quantifying the relationships in the form of distances and shortest paths and finally, c) both approaches aggregate the chosen routes into a single measure capturing the overall movement potential at a given location.

In the following, we discuss the individual parts of the Space Syntax model and highlight its specificities when compared to traditional traffic planning. Even though the methodological framework of Space Syntax covers different levels of resolution from buildings to whole cities, in this review, we discuss the neighborhood or city scale models only.

Network Model

To quantify the configurational properties of urban form, the CUM scholars represented it as a graph with the axes of movement (i.e., axial lines) as nodes and their mutual connections as edges. These axes of movement, also called axial lines, are based on the behavioral assumption that people prefer to move along straight lines.

However, the axial line also introduced several difficulties, most notably that a) these are not readily available and have to be often drawn manually (Peponis et al., 2003), b) their definition is ambiguous (Ratti, 2004), c) representing curved streets results into a large

number of small axes even though these might be perceived as one continuous element (Dalton, 2001) and d) long axes compress highly different environment into one single unit of analysis (Pafka et al., 2020).

As a reaction to this criticism, the CUM network model has been extended by two additional members, the segment and center road line map (Figure 10). The segment map is an axial line map divided into segments at their intersections (Turner, 2001). The road centerline map is the standard representation of the street network in the field of geography. The former tries to keep the idea of movement along straight lines and mitigate the problem caused by the aggregation of long sections into one element. The latter tries to utilize well established and widespread geographical standards to speed up the data acquisition process. The empirical studies reveal the different abilities of each representation to explain the movement, with the segment map being most successful in doing so. It is rapidly gaining popularity among the CUM research community (Hillier & Iida, 2005; Turner & Dalton, 2005; Varoudis et al., 2013) and has become the new golden standard in the street network representation.

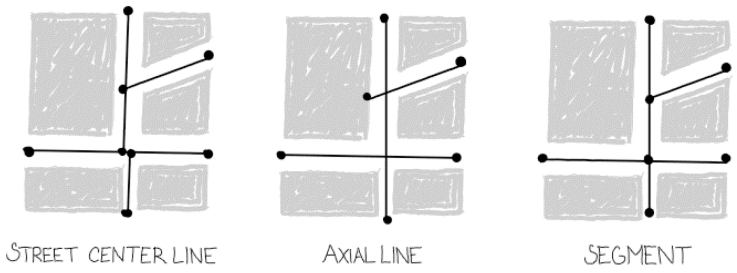


Figure 10. Network model representations in CUM.

Finally, we must emphasize that Space Syntax scholars reduce urban form to the movement system only. Plots and buildings as traditional constituents of morphological analysis and origins or destinations of movements are not explicitly represented in the model. Instead, Hillier et al. (1993) assume that the street network generates "natural movement potential," which in turn drives the allocation of the origins and destinations of the movements. This *natural movement* is the movement we would observe under the condition of an equally loaded network, i.e., the buildings and the activities are distributed equally. It is assumed that if the origins and destinations of movement are not distributed equally, these disruptions are expected to get corrected through time, and the "relation between grid structure and movement is retained" (Hillier, 1999, p9). Consequently, these corrected distributions are expected to produce the same movement patterns as the underlying *natural movement*, and thus there is no need to model them explicitly.

On the one hand, these assumptions allow a significant reduction of the CUM model, which is not only welcomed by practitioners who have to deal with limited resources. More

importantly, it can be applied in situations such as early design stages when the information about buildings, plots, and their use does simply not exist yet.

On the other hand, there has been an ongoing discussion about the validity of these assumptions as, until now, there is only little empirical evidence standing on its support (Netto, 2016; Ratti, 2004; Soja, 2001). In Bielik et al. (2017), we tested these assumptions using computational simulation and found that the *natural movement potential* assumption holds true under specific conditions, mostly depending on the movement radius. In specific, for motorized movement, the assumptions were fully proven true; however, this was not the case for the pedestrian movement. In other words, the results suggest that when it comes to origins and destinations of pedestrian movement correction force of the natural movement is limited, and their distribution should be modeled explicitly.

Route Choice Model

The quantification of spatial relationships is also based on the idea of calculating the shortest path between any origin-destination pair. As already discussed, the main question is how distance should be defined. Along with axial and segment map as an alternative representation of space, the concept of cognitive distance is another major contribution of the CUM model. It is based on developments in cognitive science suggesting that how we navigate and perceive space is based on more than metric distance and has much to do with directional changes in our routes (Golledge, 1995; Hochmair & Frank, 2002). Based on the type of network representation, CUM embraces two versions of the cognitive distance. The early development of the axial map came together with the topological distance and was later extended by angular distance used for segment and street-center line maps (Dalton, 2001; Penn & Dalton, 1994; Turner, 2001). The topological distance is measuring the minimal number of steps or axial lines required to pass in order to get from one location to another. For reasons discussed earlier, the axial map has been extended by introducing the segment map accompanied by angular shortest paths. Similarly, as the segment map increases the resolution of the axial map, the angular distance brings more detail to the topological relationships between neighboring segments. Segments lying on the same axis are assigned zero distance, which preserves the concept of axiality. However, if segments do not follow the same axis, the cognitive effort needed to navigate them is expressed as a function of their angle. This approach addresses the problem of curvilinearity in the topological analysis, where each segment of the curve is attributed to the same cognitive distance as if it would be 90 degrees turn at a crossroad. As a result, the topological distance unproportionally penalizes curved roads often composed of many small segments, whereby the angular distance remains indifferent to the number of segments as the cognitive load remains proportional to their angular deviation.

What remains an unresolved topic of discussion is the degree to which the cognitive distance on its own is able to capture human route choice behavior. The critics point out the so-

called Manhattan problem where the complete reliance on the angular or topological shortest path results in unrealistic travel patterns (Ratti, 2004). This is the result of the rectangular grid with almost everything being only two steps apart, which renders destination around the corner as equally far away as the one being many kilometers away. This problem might be less eminent in the context of organically grown European cities. However, this discussion remains open until more empirical evidence shed light on the topic.

Centrality Model – Movement Potential

The configurational properties of urban form are what CUM researchers define as the movement potential. Each element of the configuration has a different relationship with the rest of the system. It is more or less central, which results in the spatial variation of the movement potential. In other words, we can observe the relative changes of the movement potential across the urban landscape (Hillier & Hanson, 1984).

Based on what aspect of movement we are interested in, there are different graph centrality measures able to quantify it. The CUM scholars differentiate between the "to" movement and the "through movement (Hillier et al., 1987). As the descriptive terms already suggest, the movement at any location can be characterized by a) the number of trips being attracted by this location and b) the number of trips passing through this location. In graph-theoretical terms, the former is defined as closeness centrality (in Space Syntax jargon known as the "integration") and the later as betweenness centrality (known as the "*choice*").

Closeness is one of the earliest network centralities, introduced by Bavelas (1950), and later refined by Beauchamp (1965). The closeness centrality of a node is defined as the inverse of distance required to reach from one node to all other nodes in a given radius along the shortest paths (Sabidussi, 1966).

$$CC_i = \frac{1}{\sum_{j,i\neq j,d_{ij}<t}^{n} d_{ij}} \tag{4}$$

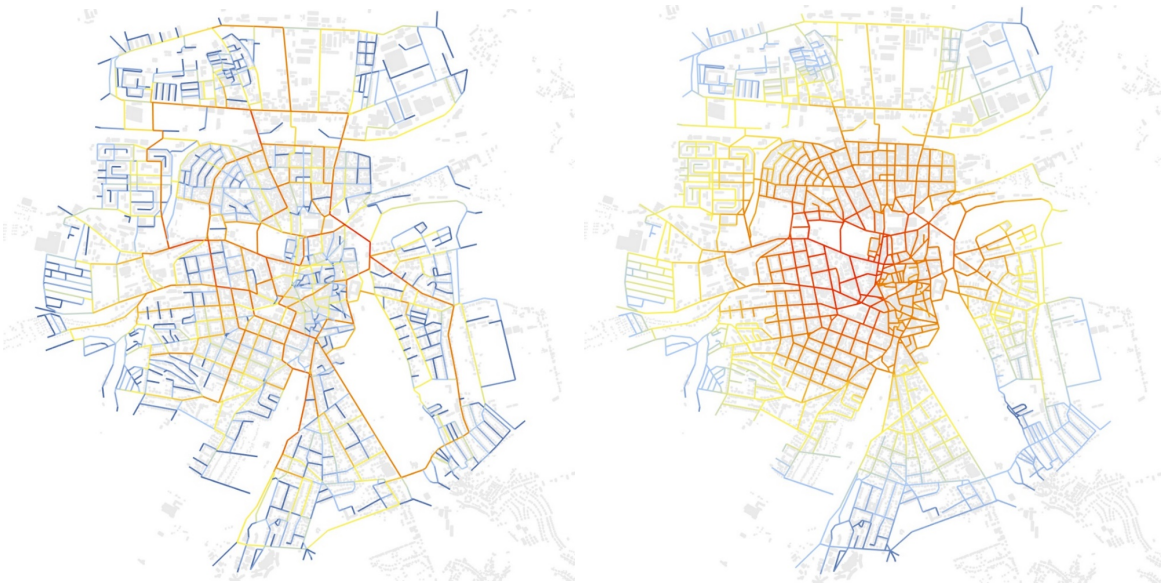
The measure of betweenness centrality was introduced by Freeman (1977) as an indicator of the importance of nodes in a social network and later adapted by geographers as a measure of flow or network loads in spatial graphs. "Considering all shortest paths in a network between all possible pairs of nodes, we can find out how often a node happens to be on a shortest path between two other nodes." (Nourian et al., 2015, p11).

$$BC_v = \sum_{i=0,\, j=0,\, i\neq j, d_{ij}<t}^{n} \sigma_{i,j}(v) \tag{5}$$

It is important to note that the betweenness centrality does not consider the flow capacity of the network, and it treats each trip as independent of the other trips in the network. As

such, it can be understood as a static movement model. The underlying assumption of unconstrained network capacity is a significant simplification and might lead to unrealistic results (Peeta & Ziliaskopoulos, 2001). The usefulness of such a static model might depend on the context, expected traffic loads, and the mode of transport.

For both, closeness and betweenness centrality, Hillier et al. (2012) proposed empirically informed normalization accounting for the effect of different system sizes. The closeness centrality is multiplied by the number of nodes NC (street segments) raised by an empirical coefficient of 1.2. The betweenness centrality is divided by the total depth of the node (i.e., distance needed to access all other nodes).



a) b)

Figure 11. Graph centrality measures a) betweenness and b) closeness visualized on Weimar's street network. (Red colors refer to higher values, blue colors to lower values).

The analysis radius is the only parameter that has to be defined for both the closeness and betweenness centrality. Similarly, as in the case of cumulative accessibility, this parameter defines the maximum acceptable travel distance and is routinely used to differentiate between the local and global type of movement as well as different modes of transport (Hillier & Hanson, 1984; Lerman et al., 2014; Raford et al., 2007). This approach, based on the fixed travel distance threshold, causes the same difficulties as the cumulative accessibility measure discussed before. Its binary nature does not reflect human travel behavior, and its value seems to be arbitrary. As a result, different researchers use different threshold values, often following the ad hoc approach and choosing whatever fits best the data. An alternative model employing distance decay function instead of a fixed threshold was proposed by Dalton & Dalton (2007). Conceptually, it is closely related to the idea of gravity-based accessibility and seems to be a promising direction for future developments.

2.3.3 Summary

The review of macroscopic pedestrian movement models reveals that even though the transportation planning (TP) and configurational urban morphology (CUM) approaches differ in how they conceptualize movement, the structure of the movement model is remarkably similar. They both embrace the idea that the proximity to destinations is the main driver of movement, which is well-aligned with the core findings of the walkability research. The approach to explaining movement is based in both cases on a) the network model capturing how the urban form is distributed and connected, b) the route choice model measuring the distances in the network, and c) the relational model capturing how everything relates to everything else.

The differences arise from the definition of each model component. The network model is in CUM approach defined by the movement infrastructure only and is represented by simple 2d lines of sight. In TP, the same representation is possible; however, it can be extended by adding buildings and plots as potential origins and destinations of movement. When it comes to the route choice model, TP tends to represent distance in metric or temporal units, whereby CUM prefers cognitive distance instead.

Finally, in CUM the relationships between all elements of the system and the resulting movement flows are assessed via the graph centrality measures, while in the TP, the same is achieved through accessibility and dynamic network loading model. Here the closer look reveals that the closeness centrality measuring the movement potential "to" any given location is in fact, just a special version of cumulative accessibility. Similarly, the betweenness centrality measuring the "through" movement potential can be seen as a simple static network loading model. We argue that the differences between the dynamic and static network loading model are in case of the pedestrian movement in low to mid-density environment negligible as the movement flows stay most of the time below the network capacity. For this reason, we consider the CUM and TP approach to network loading modeling as equivalent, with CUM offering reasonable simplification over the TP.

From a certain perspective, it almost appears as if both approaches were just identical twins with different names. Nevertheless, one substantial difference – the origins and destinations of movement highlights that despite all similarities, each of them has a different point of departure, coming from different fields of research and following different aims. The CUM model treats the origins and destinations (OD) as being equally distributed across the entire network. As a result of the constant intensity, the OD can be ignored, and movement is directly derived from the urban form.

On the other hand, the TP model picks up the variation in the spatial distribution of origins and destinations by assigning different weights to the network model. Factors such as

pedestrian movement or zoning rules create the variation in the activity pattern. As a result, the OD intensity is expected to change from one location to another (Figure 12).

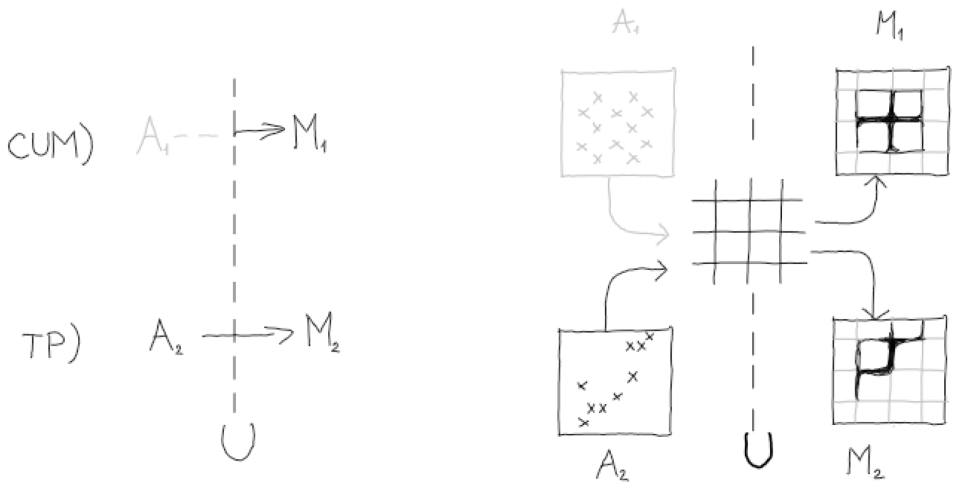


Figure 12. Schematic difference between CUM and TP approach to represent activities at origin and destination of movement.

Additionally, the CUM and TP differ in how they deal with the concept of distance. The TP uses the traditional metric or temporal distance measures, while the CUM scholars emphasize the cognitive aspects of distance.

2.4 Allocation of Activities

The allocation of human activities is central to urban planning. It is responsible for creating the physical environment in which it takes place and the rules by which it operates. It is important to realize that planners and policymakers create potential, but do not enforce individual location choices. In other words, the potentials "do not ensure that development will happen, but they do help to ensure that, if it does happen, it is designed in such a way as to be supportive" (Handy, 2009, p11). Consequently, successful planning practice is closely tied to the evaluation of the proposed potentials telling us if the plans are in line with our aims and goals.

To evaluate these potentials, different scholars at the intersection of geography, urban morphology, and economics study how and why human activities allocate in space. The general discussion has been traditionally dominated by economists (Zhang, 2002); however, in recent years, we could observe the growing interest of urban morphologists in the topic. In the following section, we review the literature on activity allocation from the urban economist (UE) and the configurational urban morphologist (CUM) perspective, pointing out the specificities and limitations of each approach.

2.4.1 Economist Approach

How do economists deal with the question of how the economy organizes its use of space? The short answer offered by Fujita and colleagues (1999) is that mostly they do not deal with the question at all. But when they do, they generally tum to a class of models pioneered in the early nineteenth century by German economist von Thünen (1826).

These models gave birth to the discipline of economic geography and urban economics as the two main domains systematically dealing with the questions of where activities take place and why (Fujita et al., 1999). The underlying location theories are based on the observation that "the location of households and firms does not exhibit a random pattern (i.e., noise), but displays many regularities" (Gorter & Nijkamp, 2001). As already observed by Adam Smith, the founding father of classical economics, all over the world, the concentration and business and population is much higher in river basin and delta areas than elsewhere (Smith, 1852). Since then, the economist proposed multiple models explaining why these patterns emerge, reacting to the limitation of previous models, and bringing new conceptual developments to the field of economics. In the following, we provide an overview of the general trends and point out the most influential ideas shaping the field. We start with the work of von Thünen, who is also widely acknowledged as the founding father of urban economics. He sought to explain the pattern of agricultural activities surrounding cities in preindustrial Germany, but the underlying ideas turned out to be universally applicable and survived until today.

Bid-Rent Curve Model

Von Thünen's abstract model is composed of an isolated town supplied by farmers in the surrounding countryside. He assumed the different yield and transportation costs of the crops to the central market. The questions addressed by the model are a) how should the land around the town be allocated to minimize the combined costs of production and transportation, and b) how will the land be allocated if there is competition among farmers and landowners? Von Thünen imagined a bidding process in which each farmer makes an offer based on the surplus he can generate and showed that competition would lead to a gradient of land rents. The maximum rent would be at the town center, and from here with the growing distance, it gradually declines to zero. Each farmer is faced with a trade-off between land rents and transportation costs with optimal allocation depending on the type of agricultural product. The resulting pattern of concentric rings of production comes into being because transportation costs and yields differ among crops. The optimal allocation of each production ring results from the "bid-rent" curves showing the relationship between the feasible rent and distance from the central market (Figure 13). Each crop is planted at a location where the farmer can afford to outbid the farmers growing other plants, or in other words, its bid-rent curve is above all other curves. Despite the striking simplicity of the model, it generates useful and unexpected insights.

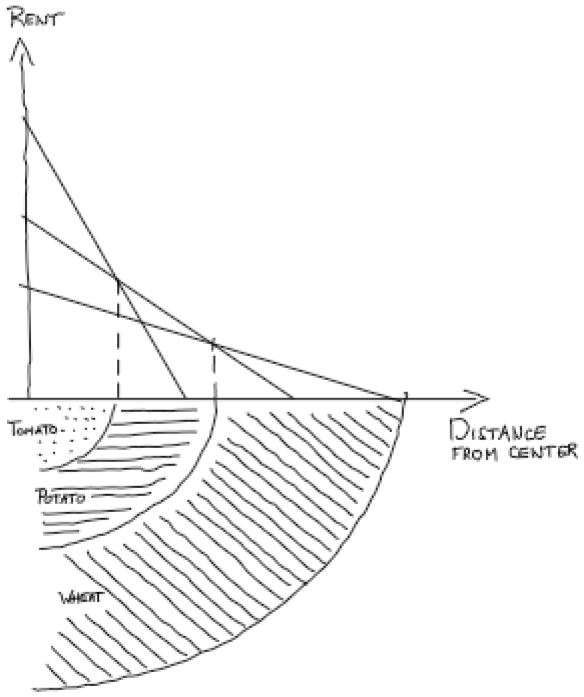


Figure 13. The land rent profile and von Thünen's rings with three crops (Adapted from: Fujita et al., 2002).

As Fujita and colleagues (2002) put it, "After all, the problem of which crops to grow where is not that easy: By allocating an acre of land near the city to someone crop, you indirectly affect the costs of delivering all other crops because you force them to be grown further away." As a result, it is a non-trivial task to estimate the outcome of unplanned market interaction. The unexpected outcome of the model is two-fold. On the one hand, it is the spontaneous emergence of the ordered and symmetric concentric ring pattern. On the other hand, it demonstrates that this unplanned outcome is not only efficient; it is indeed the optimal plan which is well in line with the classical notion of the "invisible hand."[1] The concept of bid-rent and the resulting spatial segregation is central to land allocation theory and was repeatedly picked up by and extended and reinterpreted over the course of the last century (Alonso, 1964; Marshall, 1925; Weber, 1909).

From the geographical or morphological perspective, the simplicity and reductionism of the spatial representation became the characteristic signature of these models. The economist behind them usually assumes a spatial environment that resembles a featureless plain, on which all land is of equal quality, ready for use without further improvements, and people and goods moving in straight lines (Sevtsuk, 2010). In the decades followed by the ingenious work of von Thünen, there has been a lively discussion about the economic consequences of

[1] Economic theory introduced by Adam Smith's seminal work "The wealth of the nations" suggests that decisions made by selfish individual maximizing only their own profit will lead to optimal outcome on the system level. The invisible hand is metaphor for marked force correcting the inefficiencies of the system.

his model, but only a little attention has been paid to the limitations posed by the ignorance of human geography in general, and urban form in specific. Instead, the economists pointed out that the von Thünen's model was limited in its assumption of the prior existence of the town itself and is only of little help when we ask how the town, or the system of towns came into being (Fujita et al., 1999).

Central Place Theory

The question of how the economies of scale [2] and transport costs interact to produce a spatial system of urban centers has been addressed by German geographer Walter Christaller (1933) and economist August Lösch (1940) in the central-place theory. They imagined a featureless plain, inhabited by an evenly spread population. Based on the established concept of economies of scale, the services and amenities needed by the population cannot be evenly spread and tend to form urban centers. These centers are evenly distributed to split the marked in a regular hexagonal pattern which is a result of the trade-off between scale economies and transportation costs (Figure 14).

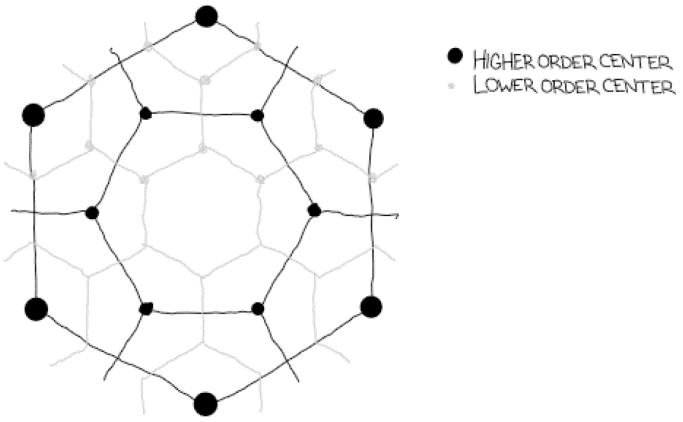


Figure 14. Overlapping market areas of hierarchical centers in the Central Place Theory.

Moreover, Christaller argued that central places form a hierarchy. This is the consequence of product diversity, price elasticity, and the existence of daily and non-daily goods. This hierarchy results in multiple layers of intertwined hexagonal pattern "where the size of the hexagons is determined by the maximum range of customers and the minimum threshold of the store" (Sevtsuk, 2010, p16). This theory was supported by a range of empirical studies determining the distribution of the central places in Southern Germany (Christaller, 1933), Estonia (Kant, 1933; Kant, 1935), and the United States (Berry, 1967) and become a landmark in allocation theory.

[2] Economies of scale is a concept based on the observation that technical, organizational and other related factors lead to the decrease in cost per unit of output as the number of produced units grow.

Nevertheless, the central place theory suffered from the same ignorance of urban form as von Thünen's model and was accused by its "looseness in reasoning" (Fujita et al., 1999, p25). Its basic idea of the efficient hexagonal lattice seems powerfully intuitive, but the theory does not describe how it might emerge. In contradiction to the established economic tradition, it gives no account of how individual actions would produce or sustain such a spatial hierarchy of urban centers (Fujita et al., 2002).

Hotellings' Linear City Model

The core assumption of the classical economic models is the rational, profit-maximizing individual. The formal study of strategic interaction between such rational individuals has been studied by scholars and gave birth to a branch of applied mathematics known as Game Theory. It draws much attention of contemporary economists because it elegantly demonstrates that the classical economic dogma of selfish behavior on the individual level resulting in an optimal outcome for the whole system, might not be true. The game-theoretical research was initialized as a reaction to the global nuclear threat of the late 1940 and gained popularity among the general public with the classic example of "prisoners dilemma" developed by Flood and Dresher (1950). It shows that two perfectly rational and purely self-interested individuals will not cooperate and suffer from the failure to do so.

This came as a surprising outcome to many since most economic models predicted just the opposite. As we already discussed by the example of von Thünen and Christaller, both models result in optimal outcomes at the system level. However, Hotelling's "linear city" model, developed 20 years before the game theory was born, could already observe the same paradox. Named after the American mathematical statistician Harold Hotelling, the model assumes a simple linear city with two vendors competing for the best location. Consumers' preferences are modeled not only as a function of their taste and the product characteristics but also as a function of the product location.

As a result, the distance to customers is a crucial parameter in profit maximization. In the example where two identical firms are placed on the line equally populated by customers, Hotelling could demonstrate that iteratively, both firms end up just next to each other in the middle of the line (Figure 15). Only at this location, the vendors reach Nash equilibrium – no further change will improve their payoff by the unilateral move. Hotelling showed that the consumers would be better off if the sellers located at the 1/4th and 3/4ths points of the line, nevertheless such a position would be not stable (Sevtsuk, 2010).

Even though the model is rather limited in its economic[3] and geographic[4] assumptions, it is considered an important milestone in the general allocation theory with a long-lasting impact on a wide range of fields, from economics political science to sociology.

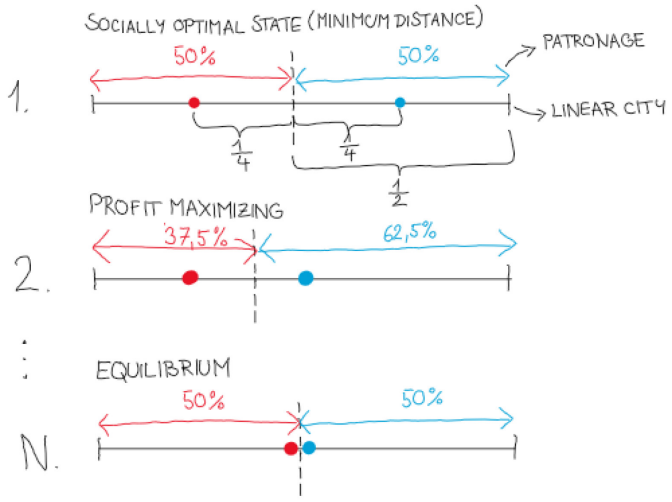


Figure 15. Hotelling's Model applied to political competition. In the course of competing for the electorate, the two parties will tend to get as close to the middle as possible to capture most of the swing voters (Adapted from Munoz-Garcia & Toro-Gonzalez, 2019).

The power of the model is in its simplicity and surprising results; nevertheless, we have to mention that if the number of competing firms in the model increases, the clustering effect disappears as demonstrated by (Lerner & Singer, 1937). DiPasquale and Wheaton (1996) illustrate a simple variation of the model with multiple retailers resulting in even distribution along the line.

Comparison shopping

It did not take long after urban economists noticed the appearance of spatial aggregation of economic activities, which could not be explained as a simple product of economies of scale. When it comes to retail and services, the models presented here assume that the competition for customers is the driving force behind the allocation choice. The result is a dispersed patronage pattern, with each customer being attracted by one retailer maximizing the total profit and minimizing the costs of service.

However, from experience, we know that similar retailers such as shoe stores are notoriously known for clustering in close proximity to each other even if the economic models suggest that they would be better of by not doing so (Bloch et al., 1991; Hise et al., 1983; Nevin &

[3] The economic competition does not allow variation in product characteristics (e.g. price, quality), and occurs only in one dimension – location. Demand is not responsive to delivered price and customers are not sensitive to transportation costs.

[4] The spatial component of the model is reduced to one dimensional line.

Houston, 1980). Eaton and Lipsey explained the clustering of homogenous commercial activities as a way to reduce customer cost for purchases led by comparison of product characteristics. Sevtsuk (2010) points out that while Hotelling regarded clustering as 'wasteful,' Eaton & Lipsey (1975) showed that it could be useful and advantageous. Their model is based on the categorization of goods and services from the perspective of the individual by comparing the probable gain from making price and quality comparisons among alternative sellers to the consumer's appraisal of the searching costs in time, money, and effort for each type of goods (Holton, 1958). As a result, the economist and transportation planners distinguish convenience, comparison, specialty, and impulse goods (Handy, 1992). Convenience goods are low-priced and frequently purchased with customers trying to reduce travel costs. Comparison goods are purchased after comparing multiple alternatives, and customers willing to travel longer distances to shop for these goods. Specialty goods are purchased infrequently with prior knowledge of what and where to buy. Impulse goods are not actively sought and do not induce additional shopping trips.

When it comes to comparison shopping, the aggregation of shops and services brings positive externalities by attracting more customers than individual retailers (Ingene, 1984). However, these positive effects are reduced by risks of bankruptcy and economic competition. If a market area can only support a certain number of retailers, the addition of another identical seller to the cluster can lead to the ruin of the whole cluster. Moreover, competition within the cluster results in lower prices and reduced profit (Cournot, 1838).

As a result, the decision to cluster is a trade-off between higher attractiveness of the whole cluster due to savings in customer search costs and lower profit margin due to the higher competition. DiPasquale and Wheaton (1996) proposed an equilibrium model for retail clustering showing under which conditions individual stores and services are expected to value competitive clusters. In general, sellers of search goods[5] have a greater proclivity toward agglomeration than sellers of convenience goods (Dudey, 1993).

Spatial Econometrics

Even though each of the economic models presented here has a wide range of variations, and thus, what we presented here is by no means a complete review of urban economics, it serves well the purpose of illustrating the general concepts and directions of the field. Overall, these models suggest that the allocation of any economic activity activities is influenced by a) type of activity, b) exogenous factors such as distance to market and customers, and c) endogenous factors, namely allocation of other activities. The allocation models differ based on which factors are considered by the modeler and how the individual economic agents (e.g., retailers, firms) interact and aggregate to the overall spatial pattern

[5] A search good is a product with features which can be easily evaluated before purchase. As a result, competing search goods can be compared.

of activities. One major difficulty of all these models is that they are largely based on deductively formulated economic theories with questionable support in empirical data (Mishra, 2007).

In the field of economics, the dissatisfaction with how the economic models were built and tested resulted in the conceptual shift giving birth to econometry. In the early 1930s, Norwegian economist Ragnar Frisch and others founded the journal *Econometrica* and argued for the application of statistical methods to economic data in order to give empirical content to economic relationships (Vernengo et al., 2020). In essence, econometrics make use of data to develop and test the economic theories. The empirical relationships in the data are estimated via methods of mathematical statistics such as multiple linear regression following the principle of unbiasedness, efficiency, and consistency (Greene, 2017).

Even though the econometrics had an enduring effect on how research in economics is done, it took another 40 years until the original methods were augmented to deal with the spatial relationships. Fujita and colleagues argue "that economist's historical unwillingness to address issues of economic geography was mainly due to the sense that these issues were technically intractable" (Fujita et al., 1999, p6). In other words, the reason why economists for a long time ignored space is the sheer inability of classical statistics to deal with it. The simple fact that in space, "everything is related to everything else" (Tobler, 1970, p234), violates the core assumption of any regression model – the independence of observations. Space brings endogeneity into the equation, which is always hard to deal with. Moreover, not only everything influences everything else, it happens simultaneously – at the same time. The approach to solving these difficulties was sketched out by the Belgian economist Jean Paelinck in the early 1970s and Paelinck and Klaassen (1979). Contrasting spatial econometrics to standard econometrics, a narrow definition is offered as dealing with "the specific spatial aspects of data and models in regional science that preclude a straightforward application of standard econometric methods" (Anselin, 1988, p8).

As it turned out, the major challenge in spatial econometrics is to model interactions without interaction data, as most studies are cross-sectional. We assume that the economic activity at any given location might be affected by the activities at the neighboring locations and vice versa. The difficulty is that we do not observe the autoregressive process composed of these simultaneous effects, but only its result. In essence, we want to estimate N^2 interactions from N observations, which is a race that cannot be won. What might be considered as a major contribution of spatial econometrics is the solution to this problem, causing so much resistance of the classic economy to deal with space. It is based on the common approach to formal modeling in general and mathematical statistics in specific when we impose structure to change the problem, which cannot be solved to another – related one which we can handle (Anselin, 2001). In spatial econometrics, the imposed structure is the matrix of spatial relationships – spatial weights matrix. This is related to the concept of the graph

and is, in essence, a formal definition of spatial relationships, neighborhoods, and distances. By introducing the spatial weights into the classic econometric regression equation, the problem of estimating N^2 interactions is reduced to the estimation of one spatial autoregressive parameter (Anselin, 2013). Several definitions of the spatial weight matrix have been proposed (e.g., contiguity or distance-based), as well as multiple types of autoregressive variables, can be modeled (e.g., spatially lagged dependent, explanatory variable, or error term). Nevertheless, what they all have in common is that the model and parameters of spatial economic interactions are estimated from data and not superimposed by the modeler.

We have to mention that even though it is relatively straightforward to identify spatial clusters and patterns of economic activities, the cross-sectional data do not provide sufficient information to identify the processes that led to the patterns (Anselin, 2010). This is also called the inverse problem, and its consequence is that with no additional information, it is impossible to distinguish between spatial dependence (i.e., interaction or true contagion) and spatial heterogeneity (i.e., "apparent contagion" or uneven distribution of environmental characteristics) (Jiang, 2015). To illustrate this issue, we can think of a town with an observed spatial pattern of disease. We see that the infections are spatially clustered, but if this is the only information we have, it is hard to tell why. The pattern could be a product of the interaction between healthy and sick individuals – "true contagion," or it could be that something else in the environment (e.g., polluted water source) makes more people at a given location sick than we would expect by random. As a reaction to the inverse problem and its consequences to establishing causality (Gibbons & Overman, 2012), specification tests have been developed that tackle both spatial dependence and spatial heterogeneity within a unified framework (Kelejian and Prucha 2007).

We conclude that the recent methodological developments and availability of a large georeferenced data set gave the spatial econometrics large momentum and marked a fast-growing field of research. Even though the current representation of spatial relationships is usually constrained to a simple topological graph (e.g., contiguity weights matrix) with growing computational power, it might be feasible to estimate model parameters within more complex distance-based spatial relationships.

2.4.2 Configurational Urban Morphology Approach

As the title of Hillier's seminal paper "Cities as movement economies" (Hillier, 1996) suggests, the relationship between urban form and urban economics has been traditionally at the very center of the CUM. The effect of location on the intensity of economic activities is a well-establish fact accepted by both the urban economist and morphologist. The saying goes, "no matter how good its offering, merchandising, or customer service, every retail company still has to contend with three critical elements of success: location, location, and location" (Taneja, 1999, p136).

Thus, the difference between CUM and UE is not "if" the location matters, but rather "how" it is represented and "what" role it plays. In CUM, the physical environment is traditionally represented at a much finer resolution, accounting for buildings and streets instead of assuming the features plain. The consequence is that distances between locations in CUM and UE models differ. What might appear close, central, and well-integrated in the birds' perspective of the featureless plain could be quite far away and segregated in the physical reality of many historically grown cities. Here, the movement is often bounded to curved paths and restricted by rivers, transportation infrastructure, or topography (Figure 16). "In fact, in built environments, the minimum distance between any two points on the streets is almost always longer than the length of a straight line between those points (Euclidean distance); let alone the extra time wasted for navigation through a complex path" (Nourian et al., 2018, p105).

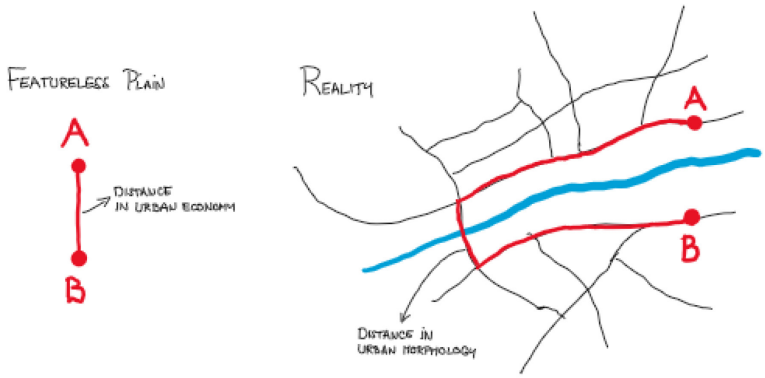
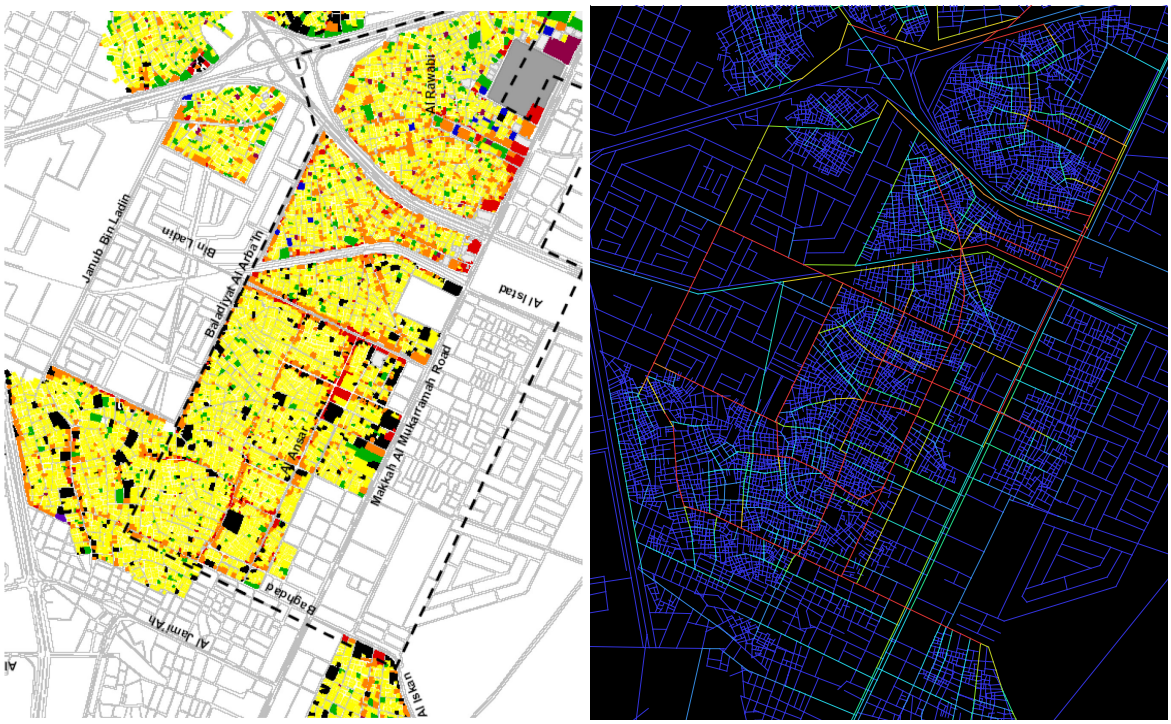


Figure 16. Difference in distances based on the representation of urban form (Adapted from Nourian et al., 2018).

The second difference is based on how the location is characterized and represented. The generally accepted idea is that activities seek locations based on how central they are or, in other words, how easily potential customers can access them. In UE, these customers are assumed to allocate at a specific stationary location (e.g., their homes or central marketplaces) from which they must travel to reach the activity. Hence, the shorter the distance, the more central is the location. This simple idea is also recognized by CUM and has been formally defined as closeness centrality (or "*integration*" in Space Syntax jargon). The difficulty is that the commercial activity in the central, well-integrated neighborhood often varies from one street to another while the closeness centrality or integration remains stable. Even one turn from a lively commercial street can completely alter the character and use of the space to a calm, low-frequency residential area (Hillier, 1998). Both streets are similarly well integrated to the rest of the city; however, their potential to attract activities is fundamentally different.

This difficulty has been identified and addressed by CUM scholars by recognizing that "place inside a city may not need to be a major trip destination but merely a pass-through nexus to generate great business opportunities" (Porta et al., 2009, p426). This straightforward observation that activities are allocating on high-frequented streets because they bring the customer right in front of their doorstep can be considered as a key contribution of CUM to allocation theory (Figure 17). The formal expression of this concept is what we described in the previous section as the betweenness centrality and can be understood as a configurational measure of flow. The CUM scholars argue that "commercial uses benefit from larger volumes of 'passing trade' and, in return, they add their own attraction"(Scoppa & Peponis 2015, p357). Chiaradia et al. (2009) and Porta et al. (2009) empirically compare several definitions of centrality and show that the through movement measured via betweenness centrality is a better predictor of commercial activity than the other alternatives.



a) b)

Figure 17. Relationship between the allocation of commercial activities and betweenness centrality in informal settlements. a) Land use distribution (yellow = residential, red & orange = commercial, green = park), b) Betweenness centrality – measure of movement flow (red = high values, blue = low values) Source: Tim Stonor (2012).

Finally, we must mention that despite the contribution of CUM to allocation theory, it remains restricted to explaining allocation through movement potential of urban form only. In general, Space Syntax scholars show only little interest in modeling economic interactions, product supply, or customer demand and seldomly differentiate between activity types in

the way the economists do. The CUM limits its explanatory framework to "functional potentials" (Hillier, 1996), which might be exploited by individuals, but remains vague about who and under which circumstances is going to benefit from it. Arguably, not every activity benefits equally from the allocation on a high-frequented street. Moreover, even those who would prefer such a location have to consider the trade-off between the benefits and higher costs. Accounting for such trade-offs is an integral part of the simplest economic models but used to be ignored by CUM.

2.4.3 Summary

When it comes to the study of how and why activities allocate in space, the urban economist and morphologist take different approaches leading to different conclusions. These differences can be traced back to the origins of each discipline, with economists focusing mostly on the market forces and interaction between seller and buyer of a product or service and morphologist looking at spatial relationships, location characteristics with a strong emphasis on the movement.

In general, UE uses to ignore the peculiarities of urban form while the CUM scholars do not pay much attention to the interaction and the resulting spillover and dispersion effect between activities. In spite of the differences between these two perspectives, we consider them as complementary and not exclusive. Their respective allocation models do not contradict but rather complement each other. Thus, we argue that combining them might be possible and highly beneficial for increasing their accuracy and reducing the potential bias.

2.5 Synthesis of the Literature

In this chapter, we started by reviewing the literature on the general concepts of model building and representations of urban form. Then we reviewed the literature on modeling the pedestrian movement and allocation of activities from the perspective of configurational urban morphology (CUM), transportation planning (TP), and urban economy (UE). In the following, we summarize the individual topics and propose a unified form-activity-movement (FAMI) interaction model combining the approaches of CUM, TP, and UE.

First, we established the idea of the model as a formal representation of the world and what we know about it. We recognized that any model is an imperfect abstraction, and strictly speaking, it is by definition wrong. Therefore, we do not care about model correctness, but instead, about the usefulness of a model. The usefulness of the model depends on its ability to answer our questions and can be formally measured in terms of its error, bias, consistency, and efficiency. As a result, when we evaluate existing models of movement and activity allocation, we do not necessarily favor more complex and precise models over the simple

ones, as this alone does not reveal how useful they are. Instead, we try to highlight the different modeling approaches and identify their limitations and contradictions (Figure 18).

As next, we defined the term urban morphology as a field of research concerned with the study of urban form. Based on the research question and the resolution of a study, different schools of urban morphology offer their own way of representing and conceptualizing the urban form. We identified the internalist-cognitive approach to urban morphology as closely related to our study's aim. Significant features of this approach, represented by Space Syntax scholars, is the focus on the relationship between the configuration of urban form and distribution of movement and activities. Known also under the term 'configuration urban morphology,' the Space Syntax school adopts a highly reductionist approach to the representation of urban form, which makes it possible to operate on the city scale efficiently. For these reasons, we adopted the CUM as the base model, which has been in the course of the literature review compared to the UE and TP approaches when it comes to modeling movement and activity allocation.

The UE tends to focus on the economic interactions between the a) activities and their patrons, and b) between the activities themselves. The resulting allocation pattern is then explained as a combination of these two forces – access to patrons and the benefits or disadvantages of spatial proximity to other activities. By comparing the UE to the CUM approach, we noticed that morphologists pay only a little attention to the interaction between activities. At the same time, economists tend to reduce the space to featureless plain ignoring basic urban elements such as streets or buildings. From a conceptual point of view, we consider the most important contribution of CUM to the allocation theory the way how a) the urban form is represented and b) how the interaction between the activities and their patrons (i.e., individuals attracted by the activity) is modeled. Regarding the former, CUM accounts for urban features such as streets and building and consider the cognitive aspects of spatial relations. Regarding the latter, in CUM, the patrons are not assumed to occupy fixed locations (e.g., homes) but are represented as dynamic movement flows. In turn, the activities which benefit from proximity to their patrons are assumed to allocate next to these flows.

As next, we reviewed the literature on pedestrian movement models and discussed the methodological differences between TP and CUM. We identified the accessibility of walking attractors as the main driver of pedestrian movement and consequently focused on the accessibility-based movement models. The main difference between the two is the role of activities as the origins and destinations of movement. The CUM scholars assume a uniform distribution of activities across space and derive movement as a direct function of the urban

form[6]. On the contrary, the TP scholars expect the activity intensity to vary across space and argue that the movement model must account for this variation.

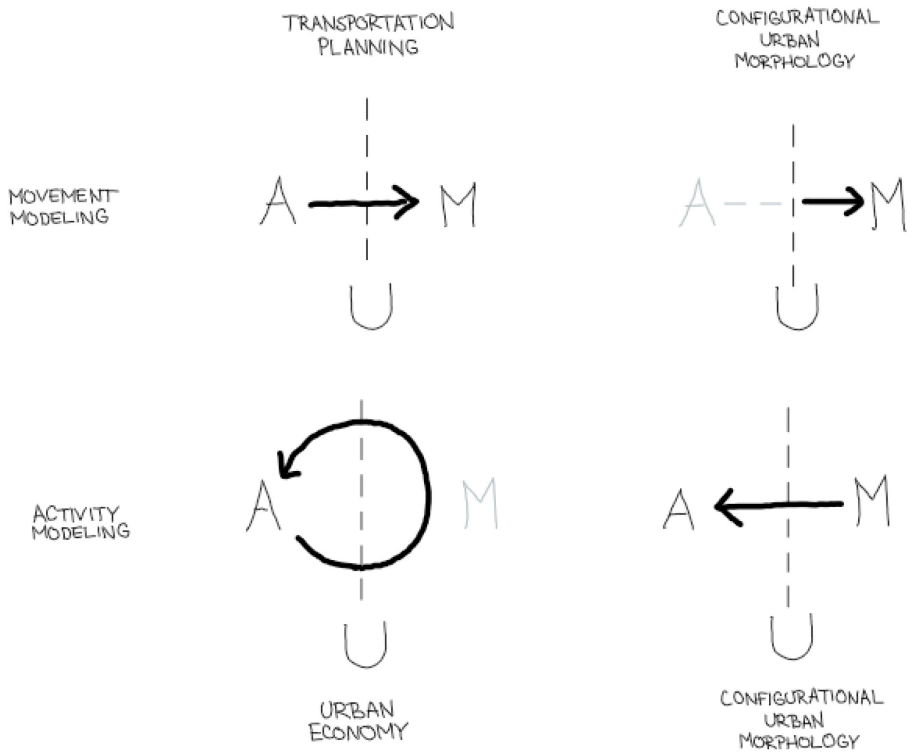


Figure 18. Schematic summary of pedestrian movement and activity allocation models. A = Activity allocation, M = pedestrian movement, U = Urban form. Arrows represent the direction of the relationship between variables.

2.5.1 Joined Form-Activity-Movement Interaction Model

In the course of the literature review, we found an ongoing discussion about the apparent contradiction between the UE, TP, and CUM approaches, raising the question of their validity. However, we argue that they can also be seen as complementary, each addressing different components of the variation in the movement and activity allocation pattern. In the following, we elaborate on this idea and present the joined FAMI model.

At first glance, the CUMs assumption about the constant distribution of activities across space is not only conflicting with the UE and TP but also with common knowledge since "urban grid is very rarely loaded in a uniform way" (Ratti, 2004, p6). We can express the

[6] The practical consequence of the uniform activity distribution is that it can be ignored when estimating movement. We can think of it as splitting the bill for dinner with a couple of friends. If everybody ordered the same, we do not need to know the price of each item to calculate how much to pay. This can be ignored as it is enough to divide the total sum by the number of persons. Similarly, the variation in the movement pattern can be calculated as a pure function of urban form, ignoring the uniform distribution of activities.

expectation of uniformity in activity distribution in the language of geo-statistics as being the result of the "random process with equal intensity across space." The argument that the randomness produces uniformity is discussed in Appendix 1, but in essence, the effect of the random process with equal intensity is like tossing a fair coin repeatedly at different locations. Eventually, we are going to get the same uniform 50:50 ratio between heads and tails everywhere.

This interpretation is useful since it allows us to rephrase the CUM approach in a less controversial manner. Instead of saying that CUM scholars assume all activities are equally distributed, we say that CUM model does not account for the whole activity pattern but only for its portion caused by chance. By doing so, we split the process driving the allocation of activities into a) homogenous component (i.e., randomness creating uniformity) and b) heterogeneous component creating variation in the activity pattern. Coming back to the coin example, the "heterogeneous component" influences the coin in a way that, at some locations, the tail is the more likely outcome, while at others, the head would come out more often. As a result, we will observe variation in the tail to head ratio across space.

One source of heterogeneity is the spatial autocorrelation between activities, as identified by UE. Another is the effect of movement on activities suggested by CUM. Additionally, there are multiple external sources of heterogeneity not considered in this study, such as zoning rules, land ownership, or the historical significance of a particular place. All these and many more are considered as additional non-morphological factors causing that urban form alone has only a partial effect on the allocation of activities and movement.

By splitting the forces driving the activity and movement pattern in the heterogeneous and homogenous components, we are able to combine the approaches of UE, TP, and CUM into a single methodological framework. Instead of seeing UE, TP, and CUM as contradictory, we conceptualize them as complementary approaches explaining different components of movement and activity allocation pattern. On the one hand, the UE and TP pick up the variation in the activity allocation pattern, while on the other hand, the CUM focuses purely on the random component of the underlying process.

The resulting joined FAMI model brings together a set of interactions suggested by the literature review on CUM, TP, and UE. It consists of the following components and their relationships (Figure 19):

1. The activity allocation pattern is composed of random (i.e., homogenous) and non-random (i.e., heterogeneous) components.
2. The two activity components attract pedestrians and generate two corresponding movement components.
3. Two movement components aggregate into the overall movement.
4. The overall movement affects the variation of the non-constant activity component.
5. The heterogeneous activity component is spatially autocorrelated (i.e., it is affected by itself).
6. The heterogeneous activity component is affected by additional non-morphological factors

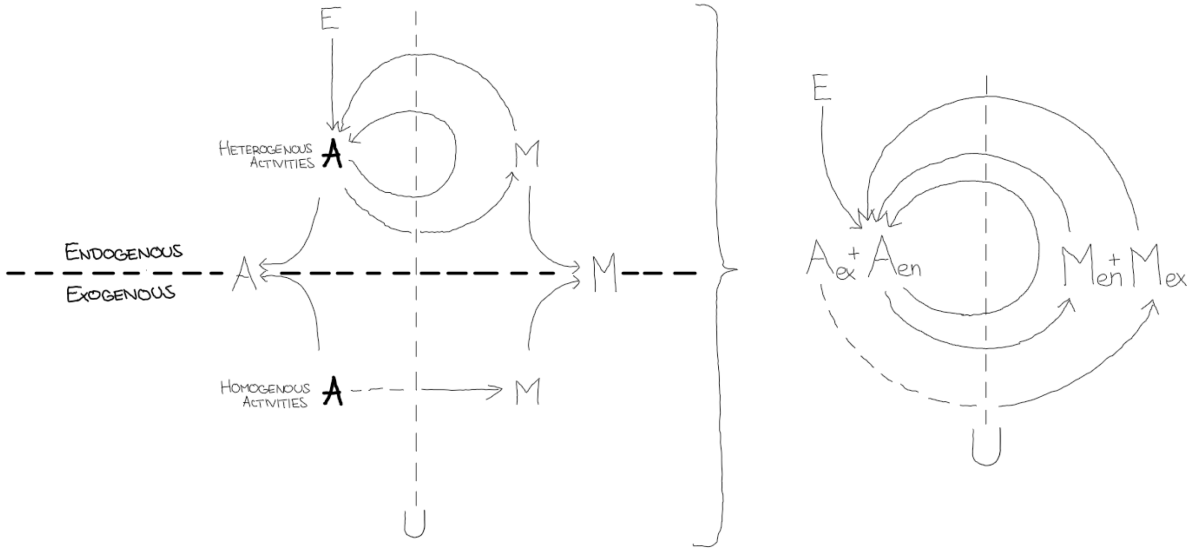


Figure 19. Exogenous and endogenous component of movement and activity allocation pattern. A = Activity allocation, M = pedestrian movement, U = Urban form.

When looking at all relationships in the FAMI model simultaneously, we observe two parallel structures. On the one hand, there is an endogenous part of the model with activity and movement components influencing each other through a feedback loop. On the other hand, we observe the exogenous part of the model with an effect going only in one direction.

The homogenous (i.e., random) activity component and the movement directly derived from the urban form are purely exogenous. It means that whatever effect they have, it never comes back to influence themselves. On the contrary, the heterogeneous activity component and the portion of movement which is not directly derived from the urban form have a bidirectional relationship. As a result, they are simultaneously influencing and being influenced by each other, which causes the endogeneity.

Consequently, we divide the model variables by their structural characteristics into endogenous and exogenous. The fundamental difference between endogenous and exogenous activity components are best visible in Figure 20. The FAMI model resembles a spiral relationship with the exogenous components standing at its beginning and the endogenous components in the middle of the feedback loop. As the uniform exogenous activity components can be ignored, the initial point of departure lies in the urban form. Together with the additional non-morphological factors not considered by this study (e.g., zoning rules, building typology), the urban form is a pivotal force giving the initial push to the activity-movement feedback loop.

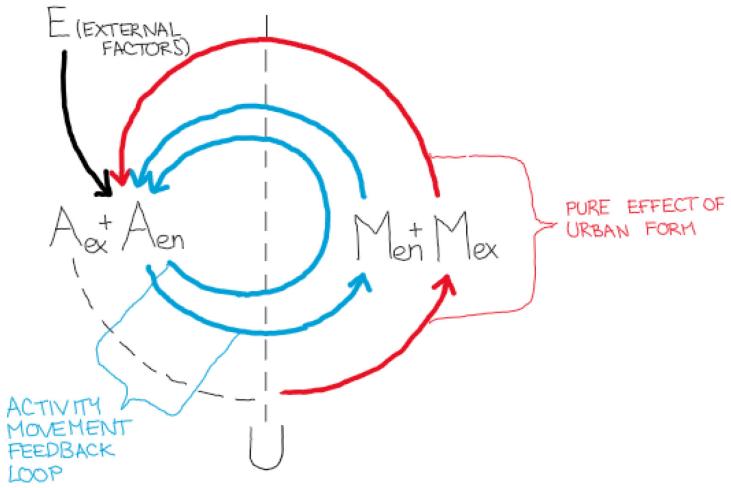


Figure 20. Joined model of interaction between activity allocation (A) and pedestrian movement (M) through the urban form (U). The model differentiates between the exogenous and endogenous components of the activity and movement pattern.

To summarize, the joined model reveals the role of individual variables going beyond the individual models proposed by CUM, TP, and UE. In CUM, the effect of FAMI is represented as a linear interaction chain starting with urban form affecting the movement, which then, in turn, affects the activities. However, the joined model suggests that the initial direct effect of urban form does not stop here but continues to propagate through the endogenous feedback loop indirectly.

This inner, endogenous feedback loop, as captured by the TP and UE, is composed of two types of interaction. On the one hand, it is the interaction between movement activities. On the other hand, it is the spatial interactions (i.e., autocorrelation) between activities themselves. Both interactions in the endogenous feedback loop are driven by a combination of two forces. It is the mixed impact of the morphological factors (i.e., the indirect effect of the urban form) and non-morphological external factors (e.g., policies, culture) not included in the FAMI model. As an illustration, we can imagine the endogenous feedback loop as an aquarium with red and blue fish interacting with each other (i.e., activity-movement

interaction). Additionally, there is a food dispenser and air pump in the aquarium, keeping all fishes alive (i.e., morphological and non-morphological factors). The important thing to realize is that the TP and UE approach tries to explain how blue fish affect the red fish and vice versa. They do not care about the individual effect of the air pump and the food dispenser.

Coming back to the joined FAMI model, this means that TP and UE do not differentiate between the indirect effect of urban form and other non-morphological factors. They entirely focus on the interaction between movement and activities. On the contrary, the CUM is focused only on the direct effect of urban form and ignores the interaction between movement and activities.

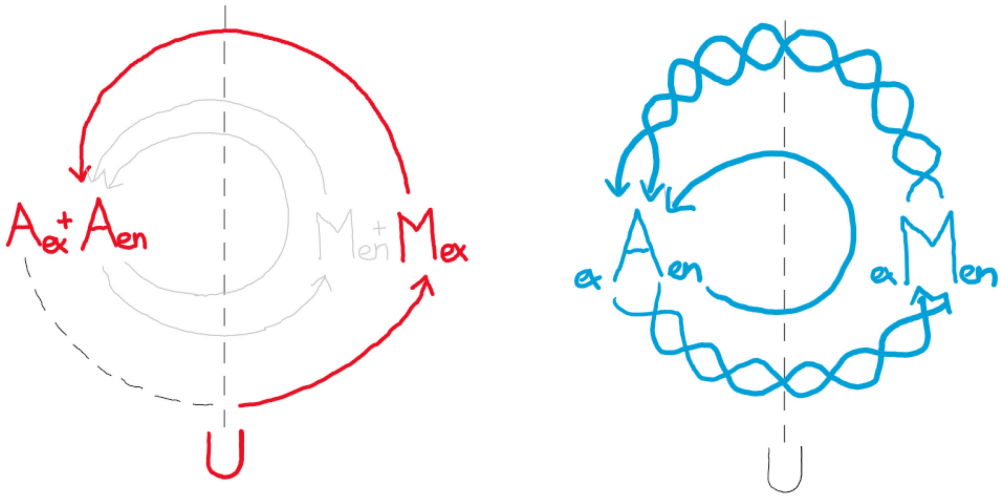
3 Research Questions and Hypotheses

The central topic of this study is the interaction between the urban form, the allocation of activities, and movement. By synthesizing the literature on CUM, TP, and UE, we found a system of interactions in which movement and activities are affected by urban form and by themselves at the same time. We differentiate between exogenous interactions going only in one direction and endogenous interactions constituting feedback loops.

When looking at the individual approaches of CUM, TP, and UE from the perspective of the joined FAMI model, we identified the following issues:

a) The CUM approach considers only the exogenous part of the interaction model and ignores the inner endogenous feedback loop between activities and movement (Figure 21a). On the one hand, it is able to estimate the pure effect of urban form on movement and activities. On the other hand, it is unclear how ignoring endogenous interactions affect the accuracy and validity of the CUM model.

b) The TP and UE approaches do not differentiate between the exogenous and endogenous part of the interaction model. They consider both at the same time when explaining the interaction between movement and activities (Figure 21b). As a result, they are unable to explain the effect of urban form, which renders both approaches of little use when it comes to informing the long-term urban planning decisions.



a) b)

Figure 21. Comparison of a) CUM approach and b) TP & UE approach from the perspective of the joined FAMI model.

We must note that the isolationist approach of the CUM, TP, and UE models does not automatically prove them right or wrong. Nevertheless, we argue that it might compromise their usefulness in the urban planning process. From this point of view, the CUM is the only approach being able to measure the direct effect of urban form on movement and allocation of activities. As such, it is of special importance for understanding the lasting impact of urban form and informing the early urban planning stages. However, we find it highly problematic that it is routinely used without any empirical evidence about the effect of ignoring substaction parts of the form-activity-movement interactions suggested by TP and UE. This is especially worrying when it comes to statistical regression models routinely used by the CUM approach when estimating the effect of urban form on movement and activities. In regression, ignoring the explanatory variable is a well-documented cause of bias (Plackett, 1949). This effect, also known under the term "omitted variable bias," occurs when a statistical model leaves out one or more relevant explanatory variables (see Appendix 2). This by no means implies that all possible effects must be considered in order to avoid the bias as this would be practically infeasible. It means, however, that under special circumstances,[7] a missing variable might cause the model to be not only less accurate but pointing in the wrong direction.

Or in other words, before applying the CUM's model, we must know the price we pay for its simplicity. Because it is often the case that "the complexity that we despise is the complexity that leads to difficulty" (Cunningham, 2004).

For the purpose of testing the validity of the CUM approach, we conduct a series of empirical tests to detect and quantify the presence of omitted variable bias. In the following, we present the associated research questions and hypotheses, which have been divided into two groups. One is concerned with interactions affecting movement. The other focuses on interactions affecting the allocation of activities.

[7] When the regression model specification omits an explanatory variable that a) is related to the dependent variable and b) is at the same time correlated with one or more of the included independent variables, the resulting model is biased (Clarke, 2005).

3.1 Pedestrian Movement

Q1: What is the effect of urban form and allocation of activities on pedestrian movement?

We estimate the direct contribution of exogenous and endogenous movement components to the overall pedestrian movement. The exogenous movement component is directly derived from the urban form and thus represents the direct effect of urban form on movement. The endogenous movement component is derived from the variation in activities (i.e., endogenous activity component). Given that the variation in activities is caused by the mixed effect of morphological (i.e., indirect effect of the urban form) and non-morphological factors (i.e., zoning rules, land ownership), we cannot clearly differentiate its source.

As a result, we describe the endogenous movement only as being a product of activity allocation without further specifying where this allocation comes from. On the other hand, in the case of the exogenous movement, we can clearly state that it comes from urban form only.

After confirming the significance of the endogenous and exogenous movement components, we test if the estimation of their contribution to the overall pedestrian movement gets biased when considered in isolation. Finally, we explore the relationship between the exogenous and endogenous movement components.

Given all these points, we propose the following research hypotheses:

H1a: Pedestrian movement is affected by urban form.

H1b: Pedestrian movement is affected by the allocation of activities.

H1c: The direct effect of urban form and the effect of activities on movement must be estimated simultaneously. Individual estimation is a source of bias.

H1d: The pedestrian movement pattern directly generated by urban form and the pedestrian movement pattern generated by the allocation of activities are significantly different.

3.2 Activity Allocation

Q2: How do urban form, pedestrian movement, and spatial autocorrelation affect the allocation of activities?

We explain the variation in activity intensity as a result of three different effects. First, we expect the activities at any location to be affected by activities at the neighboring locations (i.e., they are spatially autocorrelated). Second, we expect the activity allocation to be affected by the exogenous movement component coming directly from the urban form. And finally, we expect the activity allocation to be affected by the endogenous movement component being the result of the allocation of activities themselves.

Additionally, we assume that all these effects depend on the activity type. Different activities might be attracted, repelled, or utterly indifferent to pedestrian movement or spatial proximity to other activities, which must be considered in the process of hypothesis testing.

Consequently, we propose the following research hypotheses:

H2a: Allocation of activities is affected by exogenous pedestrian movement generated by the configuration of urban form.

H2b: Allocation of activities is affected by the endogenous pedestrian movement generated by themselves.

H2c: Activities are spatially autocorrelated. This means that activities at a given location are affected by activities at the neighboring locations.

H2d: Ignoring any of the effects described in H1a, H1b, and H1c cause omitted variable bias.

3.3 Scope of the Study

Modeling is, to a large degree, about reducing reality in a way that provides the modeler with useful insights and extends the knowledge about the phenomena in question. For this reason, modeling is necessarily about constraints and limits. This study of the relationship between urban form, pedestrian movement, and allocation of activities is constrained to a specific level of resolution and set of variables being considered.

When it comes to urban form, we investigate these relationships at the scale of whole cities, which determine how the urban form is represented. At this resolution, we limit the represent the city in the LoD1, with streets as line segments characterized by their length and relationships and buildings as extruded volumes characterized by their floor number and floor area. In line with the morphological tradition, we limit our study to ordinary buildings as these compose the majority of the built environment.

With regard to movement, the focus of this study lies on walking. We argue that even though the social, medical, and environmental benefits of walking are without dispute, the studies on urban mobility are traditionally heavily car-oriented with some emphasis on public transport. Therefore, to promote and plan for walkable urban environments, more knowledge on how and why people move is urgently needed.

Finally, we must point out that due to the scale of this study and data availability, all human interactions, needs, and decision-making are represented on the aggregated level. This does not necessarily have to reflect the choices of individuals but is still considered as useful in the context of urban planning and policymaking.

3.4 Expected Outcome

In this study, we present a joined FAMI model bringing together approaches from UE, TP, and CUM. It simultaneously accounts for the direct impact of a) urban form on activities and movement, b) the mutual interaction between movement and activities, and c) the impact of activities on themselves. Consequently, we are able to test the validity of the CUM approach by comparing its individual estimates to the outcome of the joined FAMI model. By doing so, we bring clarity about the advantages and limitations of the individual and joined models when it comes to estimating the effects of long-term planning decisions on urban form.

4 Research Methods and Data

In the following, we discuss the methods and data used to answer the research questions and test the research hypotheses. The form-activity-movement interaction (FAMI) model discussed in this section is based on the synthesis of literature configurational urban morphology (CUM), transportation planning (TP), and urban economy (UE). We start by extending the FAMI model to account for the effect of activity type. Then, we discuss the design of the empirical study in terms of data and methods used to test the research hypotheses. We conclude this section by discussing the methodological limitations posed by practical and conceptual constraints.

The major methodological challenge comes from the complexity of the interaction model. Combining approaches from three distinct fields results in an increased number of variables and relationships that must be considered. Moreover, some of the variables are practically or conceptually difficult to measure (e.g., endogenous and exogenous activity allocation), which further complicates the model estimation. As a result, the methodology presented in this chapter resembles an intricate puzzle, with a large number of missing pieces, which must be one by one estimated in order to reveal the final image. The methods used in the course of this study are largely based on linear algebra and mathematical statistics. To convey the underlying concepts to the reader with limited familiarity with these fields of study, we accompany the formal explanation with schematic graphics and illustrative examples.

4.1 The Multi-Activity-Type Interaction Model

As discussed in the synthesis of the literature, the allocation pattern of each activity type A_T can be divided into endogenous $A_{T(en)}$ and exogenous $A_{T(ex)}$ components. Consequently, each activity component of each activity type generates either endogenous $M_{T(en)}$ or exogenous $M_{T(ex)}$ movement. These two movement components constitute the total movement generated by each activity M_T .

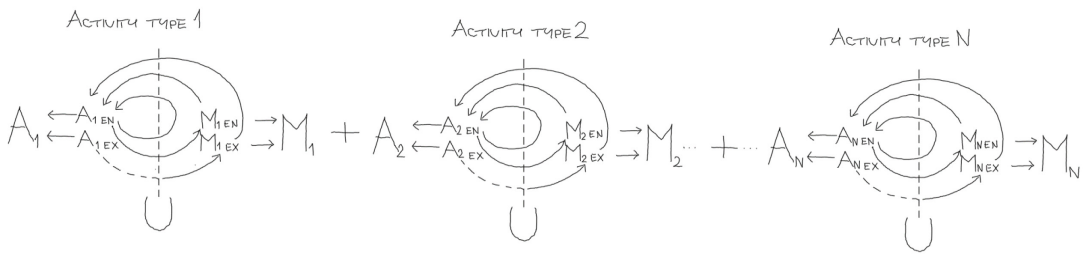
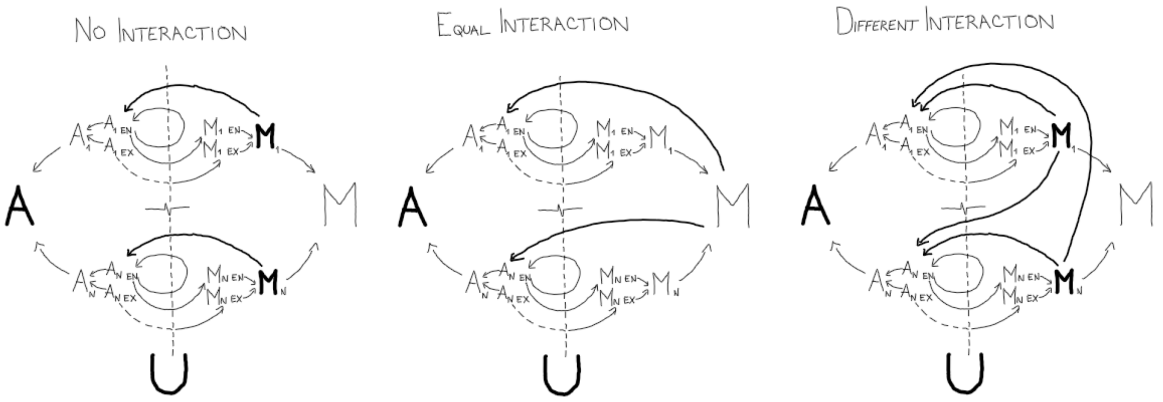


Figure 22. FAMI model for multiple activity types. A = Activity allocation, M = pedestrian movement, U = Urban form (acting as an interface of interaction between A and M.

Introducing multiple activity types brings up the question of how their respective models interact. From the conceptual point of view, we might say that the interactions between the different activities and movement types can be broadly modeled in three distinct ways (Figure 23).

a) Activities and movement types do not interact (e.g., people walking to work or restaurant do not influence the allocation of retail, only pedestrians who are on shopping trips influence the allocation of retail),
b) All types of activities and movement interact equally (e.g., people walking to work, restaurant, and shops have the same impact on the allocation of retail)
c) Activities and movement interact differently based on their type (e.g., people walking to work influence the allocation of retail differently than people walking to a restaurant)



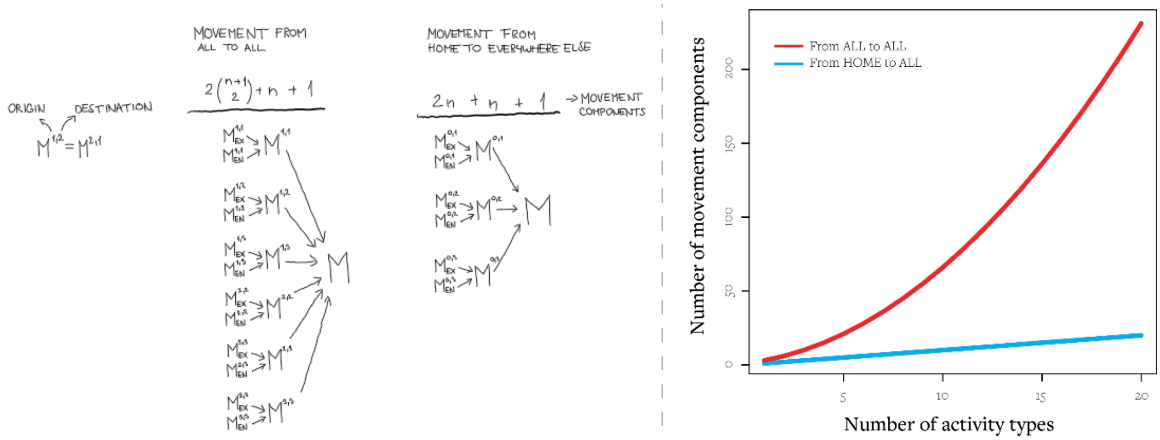
a) b) c)

Figure 23. Representations of the interaction between movement and activities based on their type (e.g., education, shopping).

As illustrated in Figure 23. each respective approach is producing a different level of complexity. With no interaction model being the simplest and the different interaction model the most complex. For practical reasons, we adopt the equal interaction model as the best trade-off between complexity and accuracy, as depicted in Figure 23b.

Another point related to multiple activity types is that each can be considered simultaneously as the origin and destination of movement. This further increases the model complexity by necessity to $2\binom{n+1}{2}$ exogenous and endogenous movement components (Figure 24a). The difficulty is that the number of model variables and relationships which must be estimated grows exponentially (Figure 24b). To mitigate this problem, we simplify the interaction model by restricting the movement origins to home locations, and only destinations follow the different activity types. According to the German national mobility

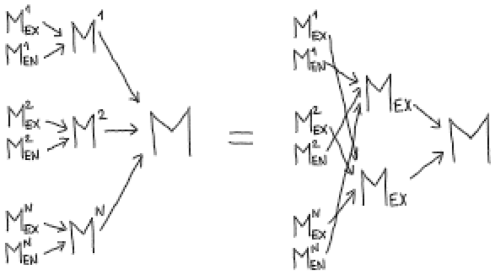
study in 2017 (MiD 2017), the home locations account for over half of all pedestrian trips in Germany and are by far the most relevant activity type when it comes to origins of movement (Appendix 7). By doing this, we arrive at a model with complexity being a linear function of the number of activity types (Figure 24b).



a) b)

Figure 24. FAMI model complexity. a) Combinatorial complexity of the model when travel from all activity types at origin and destination is considered as opposed to the simplified model considering only the travel from one origin (accommodation) to all possible activity types at the destination. b) Plot showing the exponential growth of the model components when all activity types at origin and destination are considered as opposed to the fixed model (i.e., only accommodation at origin).

To further reduce the number of variables in the FAMI model, we restructure the hierarchical relationship of individual movement components and activity types (Figure 25). Instead of combining movement components by activity type, we combine them by their structural properties (i.e., exogeneity or endogeneity).



a) b)

Figure 25. Two options for combining endogenous and exogenous movement components: a) combining by activity type, b) combining by endogeneity.

By doing so, we arrive at the total exogenous (i.e., derived directly from the urban form) and endogenous (i.e., derived from allocation of activities) movement. Both movement components are then used to explain the allocation of endogenous activities. The allocation pattern of each activity type is estimated as a function of three dependent variables: a) total exogenous movement, b) total endogenous movement, and c) spatial autocorrelation in the activity pattern (Figure 26).

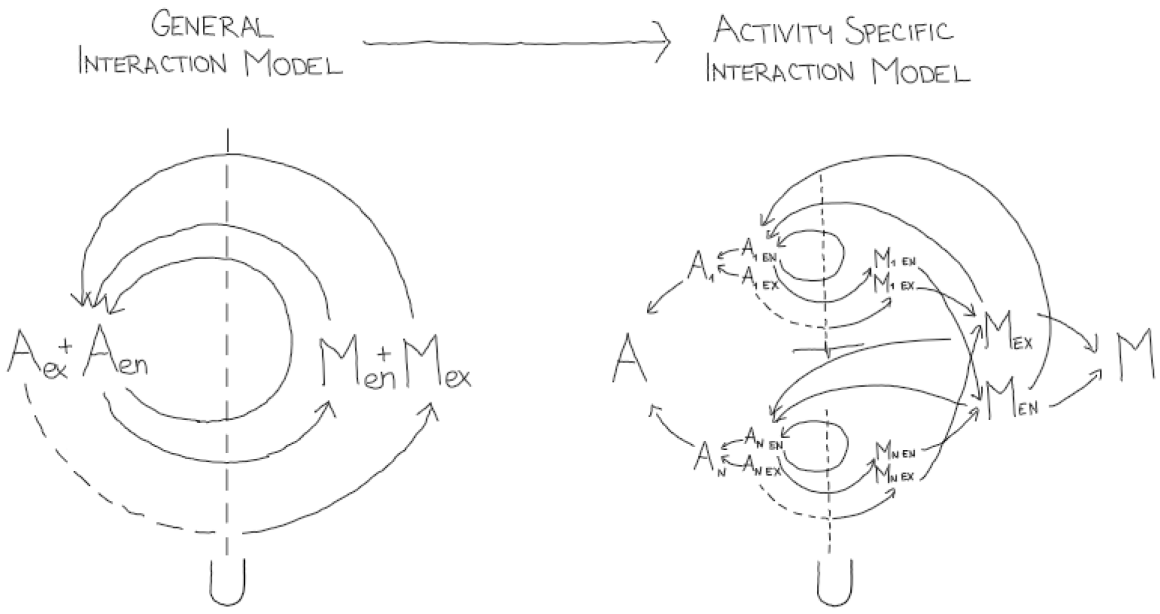


Figure 26. Comparison of the general and activity-specific FAMI model.

We must note that ideally, all variables in the model would be empirically known, and we would only estimate their relationships (i.e., the arrows in Figure 26) to answer the research hypothesis. However, the major challenge of this study is that a large portion of the variables cannot be empirically observed. The reason for this is that despite the theoretical plausibility of splitting the model variables into their exogenous and endogenous components, there is no established way how to measure them empirically. Since the endogeneity and exogeneity are properties of the whole pattern and not an individual observation, there is no such thing as an exogenous pedestrian or endogenous shop owner.

In the following, we discuss the methodological approach to collecting and estimating the model variables and the practical and conceptual limitations of the empirical data availability, and the methodological approach to overcome them.

4.2 Collecting Data and Estimating Model Variables

The empirical study was conducted in the administrative boundaries of Weimar - a historical, mid-size city located in the German state of Thuringia. Its urban form offers a data sample consisting of a wide range of morphological types, from the organically evolved medieval city center to regular grids of nineteenth-century city expansion areas and large slab-housing estates built in the 1970s (Figure 27). Furthermore, the size of the city - 64855 inhabitants on 84,420 km^2 (Statistisches Jahruch, 2018) makes it possible to cover and analyze the city as whole, which eliminates the 'edge effect' that can bias the partial analysis of larger urban systems (Gil, 2015). [8]

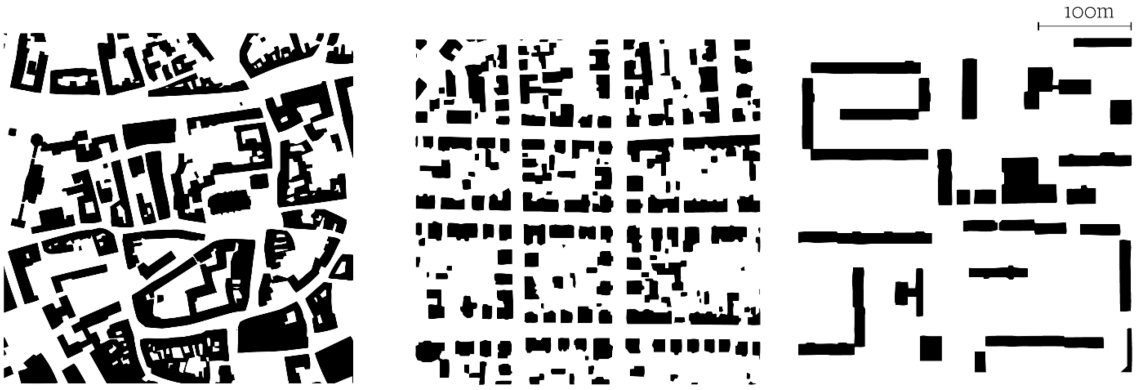


Figure 27. Street network patterns and building densities found in Weimar showing a) Historical center b) Regular grid c) Large housing estates.

[8] In the analysis of street networks, the 'edge effect' describes a bias in the analysis results as a product of portion of the network included in the analysis – the edge (Okabe and Sugihara 2012). Different measures have different degrees of sensitivity towards the 'edge effect', mostly depending on the radius of the analysis (Gil 2015). In the case study presented in this paper, we address the 'edge effect' by analyzing the whole city of Weimar. Since no additional settlements were found within the distance decay threshold (2000 m) from the edge of the city, there would be no significant change in analysis results if the edge were extended.

d) e)

Figure 27. Map of study area displaying the distribution of d) building stock and e) the street network.

It is also important to mention that Weimar is a city with far-reaching historical tradition dating back to the first Germanic settlements in 3rd century A.D. with centuries of continuous evolution, which brought it to its current form. The mature character of Weimar's urban tissue[9] is relevant since some of the statistical estimation methods adopted in this study are based on the equilibrium assumption, requiring a large degree of stability in terms of urban form, movement, and allocation of activities.

4.2.1 Urban Form

The data on Weimar's building stock and the street network was acquired from the governmental open geo-data portal administered by the state of Thuringia[10]. The dataset contains up-to-date, high-fidelity geospatial data in the LoD1 resolution with building objects represented as extruded footprints containing information on the function, built area, floor area, and building height (Appendix 4). The same data source was used to collect data on the street network geometry, which was automatically cleaned, simplified, and converted to a representation suitable for pedestrian movement modeling (Appendix 8).

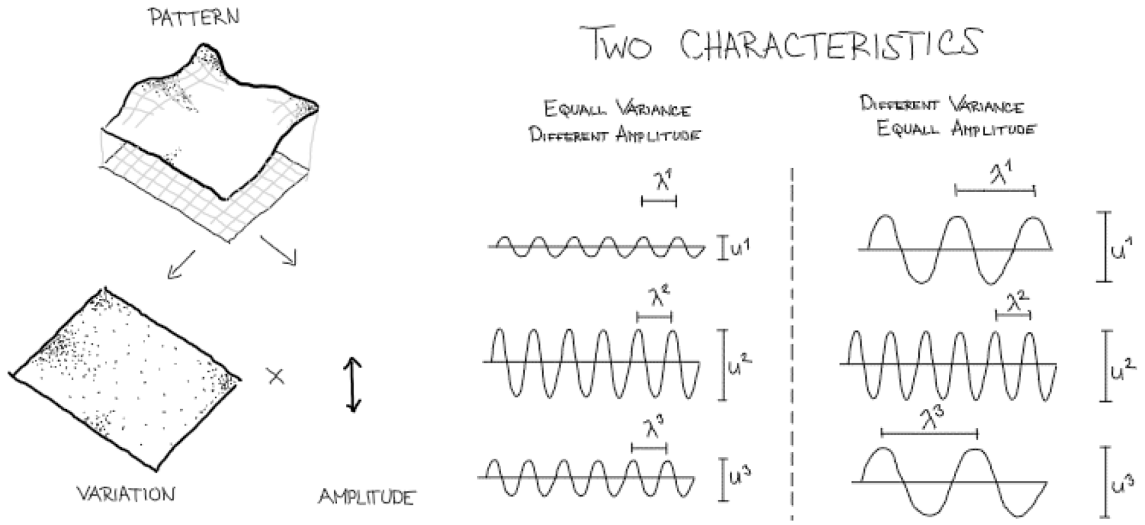
[9] The last two large scale development project in the city were the construction of condominium housing estate in Weimar Nord (1962) and Weimar West (1978). Between 1945 and 2017, the number of inhabitants kept steady at approximately 64 000 (63 976 in 1945 – 64 855 in 2017).

[10] Data is administered by the Landesamt für Bodenmanagement (https://www.geoportal-th.de/de-de/).

4.2.2 Behavior

When characterizing Weimar's pedestrian movement and allocation of activities, we describe their spatial pattern in terms of a) spatial variation (i.e., the relative difference between locations) and b) amplitude (i.e., intensity) (Figure 28). We explicitly differentiate between these two characteristics as some variables can be empirically captured only in terms of their variation, but not their intensity or vice versa. As an illustration, we can think of the spatial pattern as an audio signal which can be characterized by its frequency (i.e., tone) and the amplitude (i.e., volume). If we, for example, collect data on different tunes in the form of musical notation, we are capturing the frequency. These tunes can be then played by any skilled musician; however, each of them might choose a different level of volume as the original sound pressure level was not captured.

The methodological challenge discussed in this section is to deduce the missing characteristics where they cannot be directly recorded.



a) b)

Figure 28. Describing the spatial pattern by its variation (vector or matrix) and amplitude (constant). Illustration based on a) 3-dimensional surface and b) the example of audio signal showing the distinction between variance λ and amplitude μ.

4.2.3 Activity Allocation

The variables describing the allocation of activities form a three-level hierarchical structure (Figure 29). The bottom of the variable pyramid is composed of the exogenous $A_{T(ex)}$ and endogenous $A_{T(en)}$ activity pattern for each activity type. By combining these, we end up having the activity types A_T which aggregate into the overall activity A pattern at the top of the pyramid.

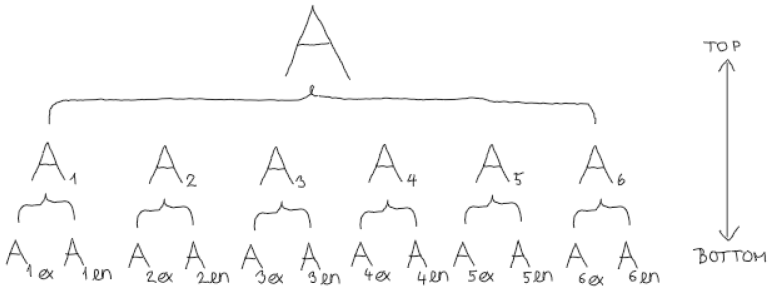


Figure 29. Activity allocation variable pyramid with overall activity pattern (A) at the top, activities by type (A_T) in the middle and activities by type and their exogenous and endogenous components ($A_{T\,(ex)}, A_{T\,(en)}$).

To describe the variation and amplitude in the overall activity pattern A and the individual activity types A_T, we combine the governmental (Geoportal-th.de) and the nongovernmental (OpenStreetMap.org) open geodata (see Appendix 5) following the travel activity categorization scheme adopted by the German ministry of transportation in the national mobility study MiD2017.

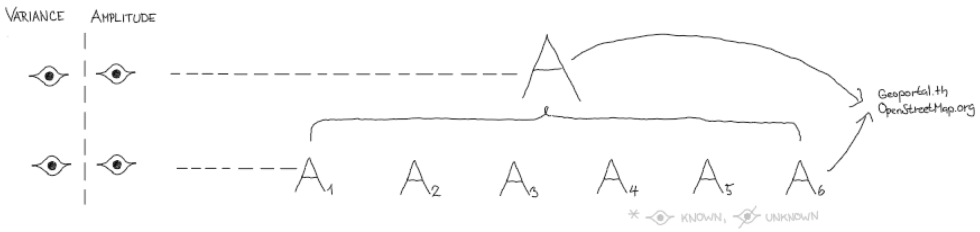


Figure 30. Empirically measuring the overall activity pattern and its components by different activity types.

The MiD2017 scheme recognizes up to 25 distinct travel activities from which 12 can be represented by the movement model adopted in this research (Appendix 5). In the case of the remaining 13 travel activities such as walking the dog or playing on the street, the destinations of travel are unknown, and thus, the resulting movement cannot be estimated [11]. The reason for the missing information on the destination of movement is that either:

a) the travel is purpose on its own, and the destination is not relevant, or it coincides with the origin (e.g., going for a walk) or
b) the travel destination cannot be localized from the activity allocation data (e.g., meeting with friends).

Such a travel behavior account, according to MiD2017 study, for 40,31% of all pedestrian trips in Weimar.

[11] As discussed in the literature review (Chapter 2.3), the pedestrian movement model adopted in this study requires the information on both, the origin and destination to estimate the resulting movement.

If all remaining 12 travel activities would be represented in the interaction model, we end up estimating 120 relationships and 64 variables[12]. Since such model estimation would be practically not feasible and hard to interpret, we reduce it to a smaller set of six most relevant travel activities, each attracting at least 3% of the total pedestrian movement. To further simplify the movement model, we restrict the considered origins to home locations only. By doing this, we were able to significantly reduce the model complexity to 56 relationships and 35 variables by accounting for 42.93% of all trips (see Appendix 5). It must be noted that this does not imply that 57% of the variance in pedestrian movement is not captured by the model, but rather that we can guarantee that the model captures at least 43%[13].

Furthermore, we test if the set of six travel activities containing a) work, b) education, c) daily shopping, d) healthcare, e) administration, and gastronomy can be further reduced by applying dimension reduction techniques (Appendix 6). We run exploratory factor analysis to detect shared variance among the six variables, which could be explained by a smaller set of underlying latent variables (e.g., shopping and gastronomy might follow the same distribution and could be explained as one category of commercial activities). Nevertheless, the results do not suggest any underlying latent structure, and we use the original set of six travel activity types in the course of this study.

The allocation of the travel activities was retrieved by merging the governmental and nongovernmental geodata. Since each data source follows a different functional categorization scheme, operates at a different spatial resolution, and has different accuracy, their combination was necessary to assure the reliability and completeness of the resulting activity allocation pattern.

By now, we covered the empirical data on the overall activity allocation pattern A and the allocation pattern by activity type A_T. However, there is no readily available data when it comes to the question of which portion of the activity allocation pattern can be attributed to the exogenous $A_{T(ex)}$ and which to endogenous $A_{T(en)}$ component of the underlying spatial process. As discussed previously, the endogeneity and exogeneity are difficult to observe on the level of individual activities and is mainly a property of the pattern. Consequently, one cannot simply collect the data on endogeneity as if it was just another property, such as the type or size of the activity.

[12] The calculation on model complexity is per one travel origin activity type.

[13] The explained variance can be higher if a) the captured movement coincide with trip chain between Home – Secondary activity – Primary activity or b) the distribution of any activity left out from the model follows the distribution of some of the included activity. In such cases, the resulting model will have higher explanatory power but might run the risk of omitted variable bias and parameter overestimation (see Appendix 2).

However, the data on the variation of $A_{T(en)}$ and $A_{T(ex)}$ can be derived from the structural relationship between these two activity components and the resulting activity pattern (Figure 31). The general idea is based on the fact that the exogenous activity component is uniformly distributed (i.e., it is constant with variation equal to zero) while the endogenous component is picking up the variation in the overall activity pattern. Thus, if the overall activity pattern A_T is known, it has the same variation as $A_{T(en)}$. Nevertheless, their amplitude remains unknown. To illustrate this principle, we can imagine an experiment in which we turn off and on the heating in the living room while measuring air temperature in the bedroom. As the warm air spread through the apartment, we might be able to deduce the temperature variation in both rooms. However, the absolute temperature (i.e., the amplitude) in the living room will remain unknown.

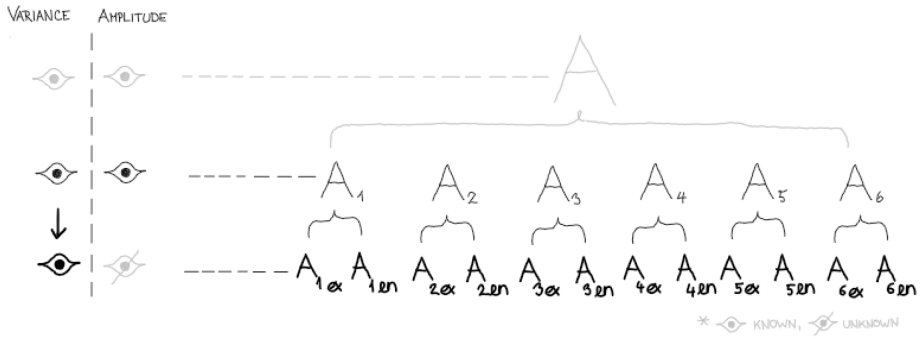


Figure 31. Estimating the variance in the exogenous and endogenous activity components.

To sum up, even though the exogenous and endogenous activity pattern components cannot be empirically observed, their variation can be derived. In the case of $A_{T(en)}$ it is the same as A_T, while in the case of $A_{T(ex)}$ the variation is equal to zero. However, it is important to realize that the intensity of the $A_{T(en)}$ and $A_{T(ex)}$ components remain unknown. This poses a significant challenge for the estimation of other variables and relationships and will be further discussed in the following.

4.2.4 Pedestrian Movement

In this section, we look at each of the 15 movement-related variables in the FAMI model and discuss if it can be either directly observed or must be derived through simulation or statistical estimation. The variables can be sorted into a hierarchical pyramid-like structure with the endogenous and exogenous movement components per each activity type $(M_{T(en)}, M_{T(ex)})$ at the bottom, aggregated endogenous and exogenous movement (M_{en}, M_{ex}) in the middle and total movement M at the top of the pyramid (Figure 32).

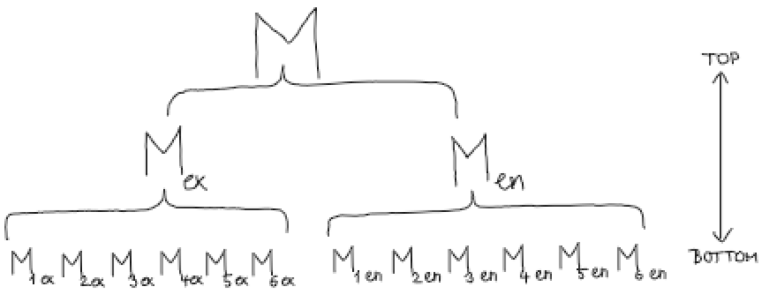


Figure 32. Movement components pyramid with overall movement pattern (M) at the top, movement by its exogenous and endogenous components (M_{ex}, M_{en}) in the middle and movement by activity type and exogenous and endogenous components ($M_{T\,(ex)}, M_{T\,(en)}$).

4.2.4.1 Defining Movement

When it comes to capturing movement, we first address the question of how to represent it. We argue that movement is a multidimensional concept that cannot be described by any single quantity and can be fully understood only by looking at multiple characteristics simultaneously. To illustrate the idea of multidimensionality, we can think of the weather. Its spatial pattern can be described in terms of temperature, air pressure, wind speed, or rainfall, just to mention a few. Based on our interest (e.g., going hiking, biking, sailing), a combination of different characteristics might be relevant.

In the following, we briefly introduce six distinct movement characteristics, out of which four are relevant for this study and are discusses in detail in Appendix 10. We show that each movement characteristic captures a different aspect of the movement while being structurally related to all other characteristics. The movement characteristics presented here are by no means the complete list of the possible ways of quantifying movement. They were chosen mainly to answer the research questions stated in this study and to demonstrate the idea of movement as a multidimensional concept.

In general, movement is characterized by its origin, destination, and the path connecting them. Thus, based on which aspect of the movement do we focus on, we can speak about the a) **From**, b) **To**, and c) **Through** movement.

In the case of the "From" movement, we characterize the movement at its origin so we can tell how different locations affect the walking behavior of those who live here. In a similar way, the "To" movement considers only pedestrians who end their trip at any given location and can be seen as a measure of movement attractivity. Finally, to investigate the effect of movement on the allocation of activities, we assume that these are influenced by the pedestrian movement passing through a given street regardless of where their trip starts or ends. Accordingly, we term it as the "Through" movement.

We quantify each movement characteristic based on how often and how far people walk. The first gives us an idea about the **Frequency** and the later about the **Volume** of the

movement. When combining both aspects together, we end up having six distinct characteristics of pedestrian movement, from which four are considered throughout this study (Table 1, Figure 33).

Table 1. Six movement characteristics. From-Frequency, From-Volume, Through-Frequency, and Through-Volume are relevant for this study.

	Frequency (Nr. Of trips/time)	Volume (km/time)
From	**From-Frequency**	**From-Volume**
Through	**Through-Frequency**	**Through-Volume**
To	To-Frequency	To-Volume

The *Through-Frequency* and *Through-Volume* are, by any means, the most relevant movement characteristics for this study as they are assumed to impact the allocation of activities. Additionally, we consider the "From-Frequency" and *From-Volume* movement characteristics to quantify the effect of home location on walking behavior (i.e., walkability) and to demonstrate the multidimensionality of movement.

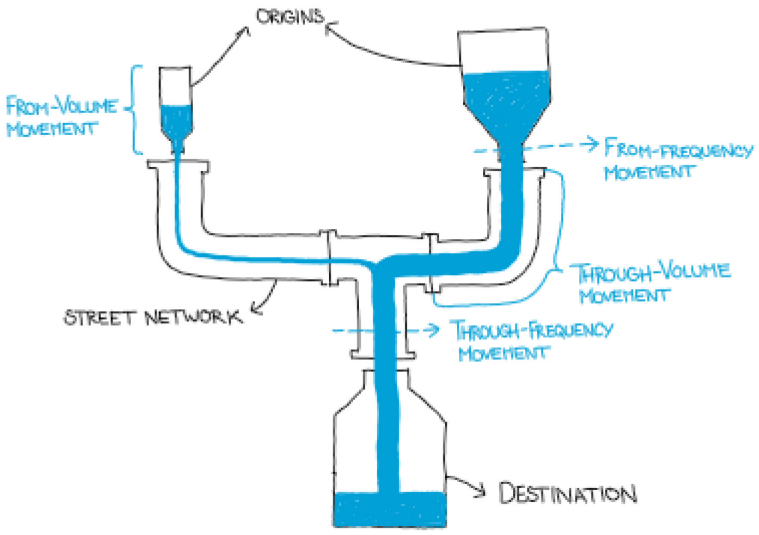


Figure 33. Illustration of the movement as a multi-dimensional concept.

4.2.4.2 Measuring and Estimating Movement

In the following, we present the data on the movement variable pyramid. For each variable, we either collected the data on all four movement characteristics or estimated its spatial pattern if empirical data was not available.

Measuring the Variation and Amplitude

Like the activity allocation, in the case of pedestrian movement, not all variables can be empirically observed, and their pattern cannot always be described in terms of its variance and amplitude (Figure 34).

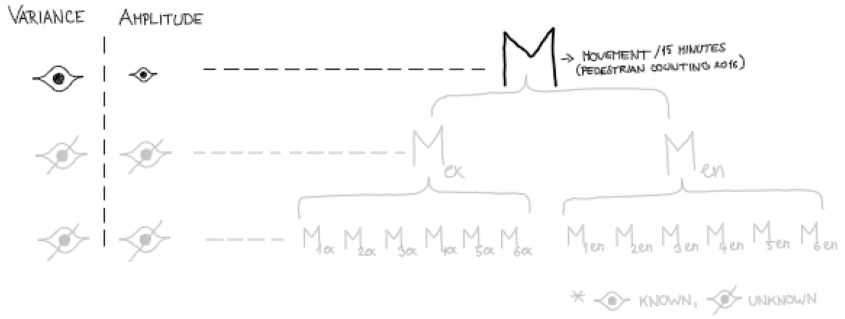


Figure 34. The variation and amplitude of total pedestrian movement (Through-Frequency) coming from pedestrian counting 2016.

We start by capturing the variance and amplitude of the total movement pattern. The empirical data on *Through-Frequency* was manually collected by the authors in 2016. We recorded the number of pedestrians passing through 100 locations spread across the study area. Each measurement took 15 minutes and was conducted on three different daytimes and three different weekdays, coming to a total of nine measurements at each location (Appendix 7). As a result, we could investigate the effect of daytime and weekday on the movement pattern and choose the best representative temporal unit of measurement. As discussed more in detail in Appendix 7, we found that on the one hand, the movement pattern is relatively stable across the weekdays, while, on the other hand, it varies largely between different daytimes. Given these results, we use a single day as a temporal unit for scaling the relative to absolute movement. In essence, it is the smallest common denominator, representing a trade-off between stability with the detail of temporal resolution.

To extrapolate the measured pedestrian frequencies from 15 minutes to full day, we derive the daily pedestrian movement frequency function for Weimar and scale each measurement accordingly (Figure 35). The function is derived from the travel diaries collected in the study area in the scope of the MiD2017 study as it keeps track of the start and end times of all journeys in the study area. In effect, we draw the daily movement distribution curve for

each of the 100 measurement locations and calculate the area under the curve accounting for the total number of passing-through pedestrians per day.

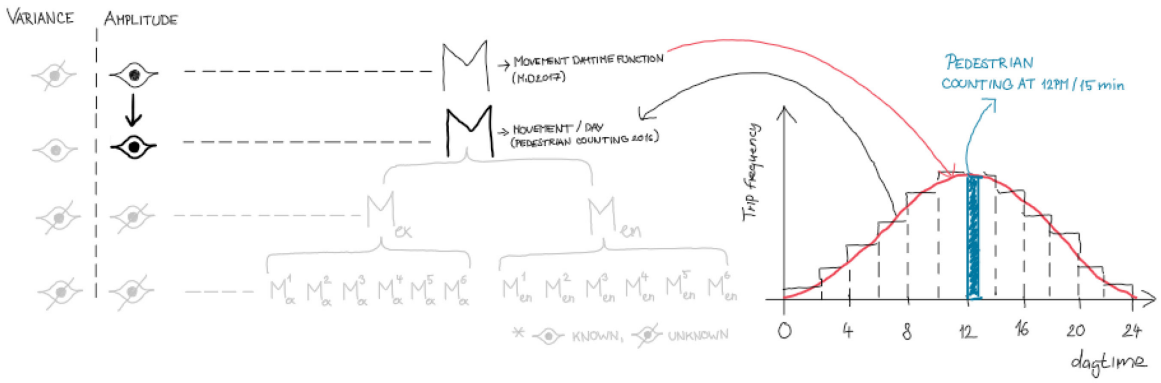


Figure 35. Scaling the amplitude of the pedestrian movement pattern (Through-Frequency) from the original 15 minutes counting interval to whole day.

Simulating the Variation

Like in the case of the activity allocation pattern, it is conceptually problematic to collect data on the endogenous and exogenous movement components. The individual pedestrians do not distinguish between walking to the exogenous or endogenous activities. They only go to work, school or shop. Even though we cannot empirically measure the exogenous and endogenous movement components, we estimate them through simulation. As discussed before, we know the variance of exogenous and endogenous activity patterns, which can be, in turn, used as input for the pedestrian movement model (Figure 36). In the case of the endogenous movement, we run one simulation model for each activity type, while in the case of the exogenous movement, all activity types have the same variance (i.e., variance equals zero), so the number of simulations can be reduced to one (Figure 36).

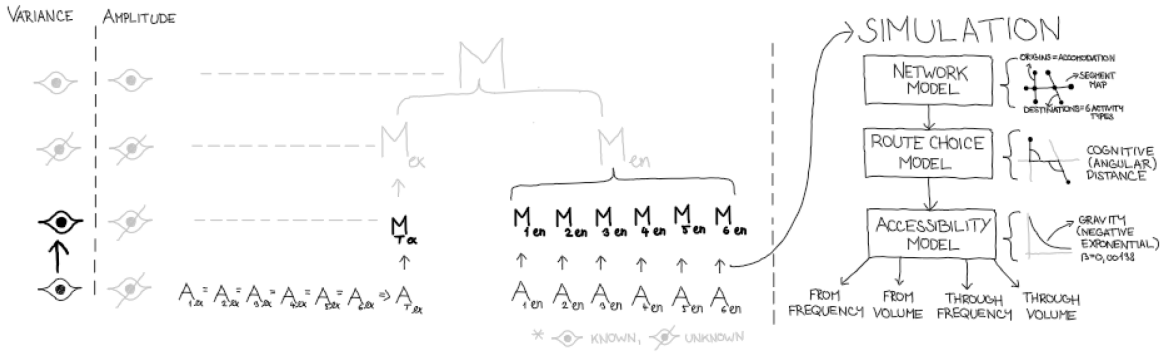


Figure 36. Simulating variance in the exogenous and endogenous movement pattern by activity type. Since exogenous activities do not differ in their variance, we can simplify the model to one generic activity variable A_{ex}^{T} and simulate the resulting generic movement M_{ex}^{T}.

The simulation model is based on the joined methodology used by TP and CUM as discussed in Chapter 2.3. We start by defining the network model, with each street segment being weighted by its respective activities. The activity weights represent the total amount of floor area (m^2) per activity, which can be found at a given street and serve as origins (i.e., accommodation) and destinations (i.e., six travel activities) of movement.

As next, we specify the parameters of the route choice model – how distance is defined and calculate the shortest paths between the origins and destinations of movement. Since the literature suggests several approaches for defining distance (e.g., metric, cognitive), we run an empirical study in Weimar to select the best performing alternative. As discussed in Appendix 7, we recorded the daily travel paths of 50 participants and compared them to the shortest metric and shortest angular paths (Figure 37). We found that in the case of Weimar, the cognitive shortest path model minimizing the angular deviation between origin and destination of travel was performing better than the metric alternative. Therefore, in the scope of this study, we adopt the minimal angular deviation as optimization criteria for path selection.

Figure 37. Route choice study 2017. a) Set of 203 unique pedestrian paths as captured by the Path selection 2017 study. Paths are laid on top of each other, with the color intensity indicating multiple paths sharing the same street segment. b) one observed path (black color) and the shortest angular (green color) and shortest metric (red color) alternatives.

Finally, we select the accessibility model capturing the willingness to travel and calibrate its parameters. These are used to simulate the spatial distribution of all four endogenous and exogenous movement characteristics for each activity type. Based on the literature review, we consider the gravity-based accessibility model as the best trade-off between complexity and accuracy. It addresses the arbitrariness of the radius threshold of the cumulative accessibility and offers relative simplicity over the utility-based accessibility models (i.e., it requires fewer data to calibrate. See Chapter 2.3).

We calibrated the distance decay function of the gravity accessibility model (Figure 38) from empirical travel data collected in the study area in the scope of the MiD2017 study (Appendix 14). We found the resulting negative exponential distance decay function (Equation (6) for Weimar to be slightly steeper than the national average[14].

$$f(d_{ij}) = \frac{1}{e^{\beta d_{ij}}} \tag{6}$$

With gradual fall reaching 50% penalty at 502m distance, 90% at 1668m and more than 99% at 3337m and more. In essence, it means that a destination loses over 99% of its ability to attract pedestrians if it is more than 3.3km far away.

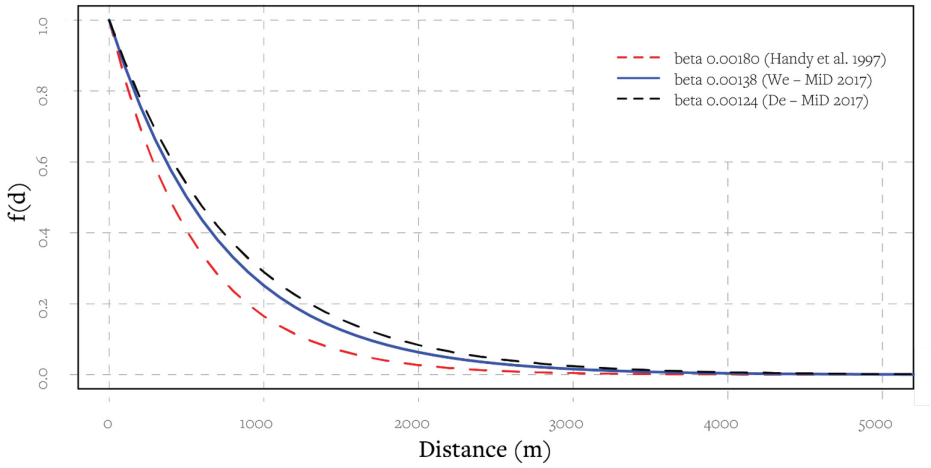


Figure 38. Calibrated negative exponential distance decay function with beta coefficient β = 0.00138. The function for Weimar (blue curve) is compared with the average for the whole of Germany (black curve) and reference function fitted in 1997 by Handy et al.

After calibrating the parameters of the pedestrian movement simulation model (i.e., gravitational distance decay function, cognitive shortest path), we estimate the endogenous and exogenous movement components of the movement variable pyramid. In specific, we

[14] Steeper curve means being less willing to walk longer distances. In other words, we found that pedestrians in Weimar walk shortest distances than the national average.

estimate the four-movement characteristics (i.e., *From-Frequency, From-Volume, Through-Frequency*, and *Through-Volume*) for each of the six activity types. We must emphasize that since the endogenous and exogenous activity components are known only in terms of their variation, the estimated movement components will also reflect only the variation, but not the amplitude of the movement pattern.

The **From-Frequency** estimates the frequency of trips starting at location v and is defined as the product of the weighting at the origin w_i and destination w_j of movement and the distance decay function $e^{\beta d_{ij}}$.

$$FromFrequency_v \sum_{i=v,j=0,i\neq j}^{n} \frac{w_i w_j}{e^{\beta d_{ij}}} \tag{7}$$

Conceptually, we see the mathematical representation as a product of two components – the travel demand and travel supply. The former is simply representing the distribution of people willing to travel (i.e., accommodation activity intensity w_i), while the latter is proportional to the attractiveness of the destination (i.e., travel activity intensity w_j) and inversely proportional to the function of the distance to the destination (i.e., negative exponential distance decay function $\frac{1}{e^{\beta d_{ij}}}$).

$$traval\ demand = w_i\ ;\ traval\ supply = \frac{w_j}{e^{\beta d_{ij}}} \tag{8}$$

Finally, we want to point out that this operationalization of the From-Frequency movement characteristic is identical to the graph centrality measure known as *Gravity* centrality (Hansen, 1959).

The estimation of the **Through-Frequency** at location v is based on the same travel supply and travel demand components as the *From-Frequency* with the difference that also journeys which do not start at v but are passing through v are considered.

$$ThroughFrequency_v \sum_{i=0,j=0,i\neq j,S_{ij}\epsilon\, S_v}^{n} \frac{w_i w_j}{e^{\beta d_{ij}}} \tag{9}$$

The *Through-Frequency* is conceptually close to graph centrality measure *betweenness* (Freeman, 1977). The only difference is that instead of the Boolean travel impedance function (i.e., distance threshold as in the cumulative centrality), we adopt a more realistic

gravity-like distance decay function as in the case of the *From-Frequency*. As a result, we call the *Through-Frequency* as *Gravitational Betweenness* centrality.

The **From-Volume** builds upon the *From-Frequency* specification with the addition of each journey being multiplied by its length d_{ij}. By doing this, we characterize the total travel distance of all journeys undertaken from the location v.

$$FromFrequency_v \sum_{i=v,j=0,i\neq j}^{n} \frac{w_i w_j}{e^{\beta d_{ij}}} d_{ij} \tag{10}$$

Similarly, the **Through-Volume** is based on the *Through-Frequency* specification with the addition of each journey being multiplied by the length d_v of the street segment at the location v.

$$ThroughVolume_v \sum_{i=0,j=0,i\neq j,S_{ij}\epsilon\, S_v}^{n} \frac{w_i w_j}{e^{\beta d_{ij}}} d_v \tag{11}$$

To summarize, the estimation of the variation of all movement characteristics is based on the same travel supply-demand model using the same travel impedance function and the cognitive shortest paths. As a result, all four characteristics are structurally related (Appendix 10). The practical implication of their relatedness is that if we find the amplitude of any single characteristic, we can directly derive the amplitude of the remaining ones (Appendix 11).

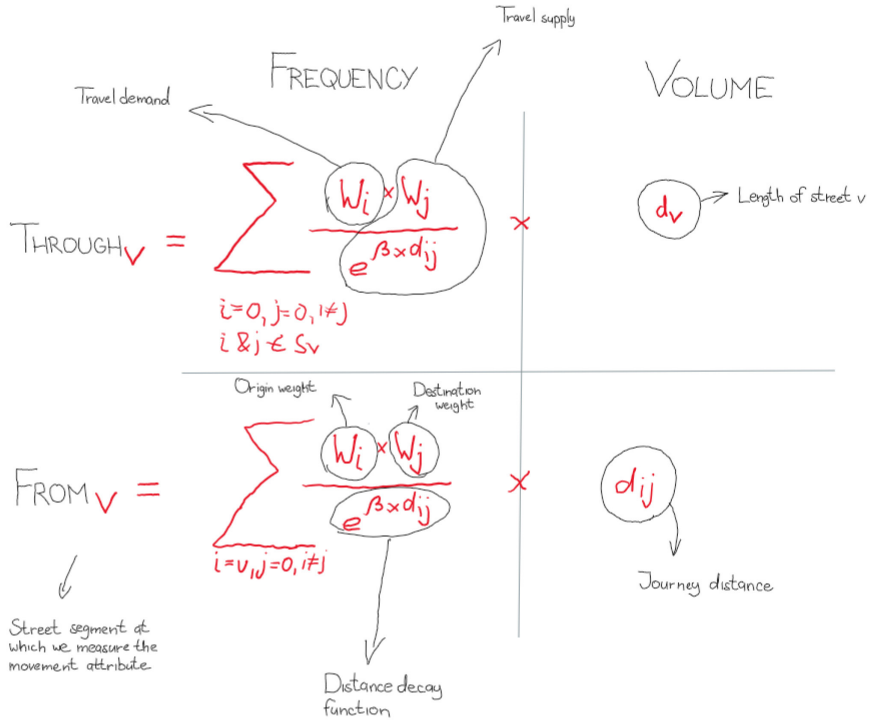


Figure 39. Relationship between the From-Frequency, From-Volume, Through-Frequency, and Through-Volume movement characteristics.

The last point to be mentioned here is the conceptual difficulty of estimating the endogenous movement $M_{T(en)}$ from allocation activity data, which captures the combined effect of $A_{T(en)}$ and $A_{T(ex)}$. We argued that even though we cannot collect the data on $A_{T(en)}$, its variation can be derived from A_T. When it comes to variation in activity pattern $A_{T(en)}$ it is equivalent to A_T. As a result, $A_{T(ex)}$ can be ignored because it does not vary (i.e., it is constant). However, when it comes to movement, even the constant activity distribution creates variation. If activities are distributed equally, the resulting movement will not be the same everywhere (Figure 40). The consequence is that $A_{T(ex)}$ can be ignored when it comes to variation in activities but not when it comes to movement. In effect, the variation in $M_{T(en)}$ estimated by our method is the combination of $M_{T(en)}$ and $M_{T(ex)}$. To arrive at the true $M_{T(en)}$ we must filter out the $M_{T(ex)}$ as described in Appendix 12.

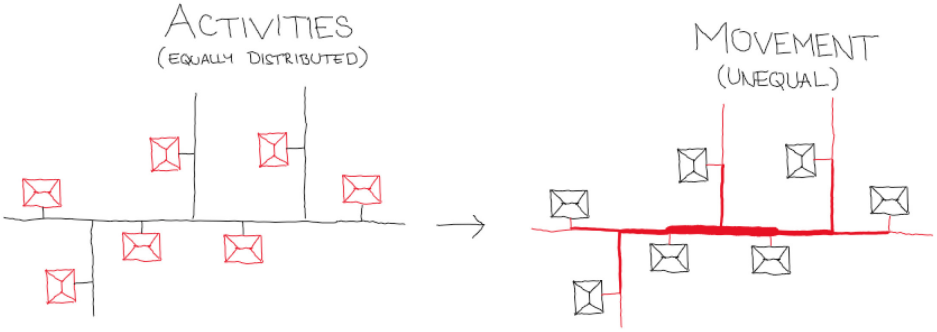


Figure 40. Homogeneous distribution of activities (i.e., zero variance) results in heterogeneity in movement flows (i.e., non-zero variance).

Estimating the Amplitude

As already discussed, the patterns of endogenous and exogenous activity components used to simulate the movement were defined only in terms of their variation, but not the amplitude. Consequently, the simulated movement is revealing only the relative variation in movement pattern but not its absolute magnitude. We know that location A generates twice as many trips to given activity than location B; however, it does not tell us how many it is in absolute terms (e.g., person/day). Hence, we cannot compare the movement intensity of different sources of movement (e.g., work vs. education) and, more importantly, combine them to get the aggregated exogenous and endogenous movement (M_{en}, M_{ex}).

We can think of the problem as having different sources of sound (i.e., soundtracks), which should be combined into a song. Each track can be described in terms of its wavelength (i.e., variance) and loudness (i.e., amplitude). The problem we face is that we are missing some of the tracks, and for the rest, we were provided only with the information about the wavelength. As depicted in Figure 41, to combine the individual tracks and figure out the missing pieces, they must be in the right relationship to each other (i.e., their amplitude must be correctly scaled). Otherwise, their combination will produce different songs. The key idea to realize is that the amplitude of track A and B does not have to be known in absolute terms (i.e., the actual acoustic pressure in of the original recordings), but only their relative proportion must be correct (Figure 42).

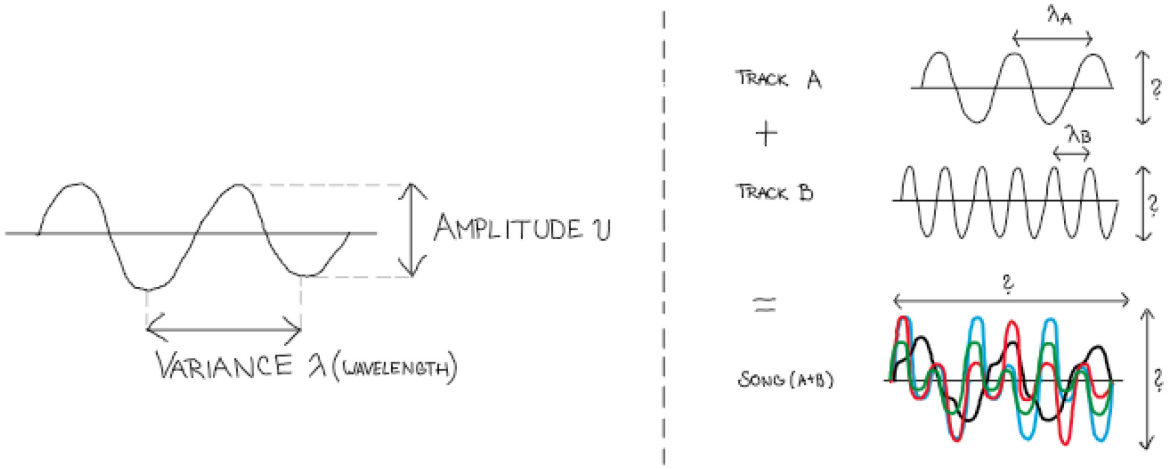


Figure 41. Combining two patterns with unknown amplitude illustrated by the example of sound waves. Without the information on the amplitude, different results are possible.

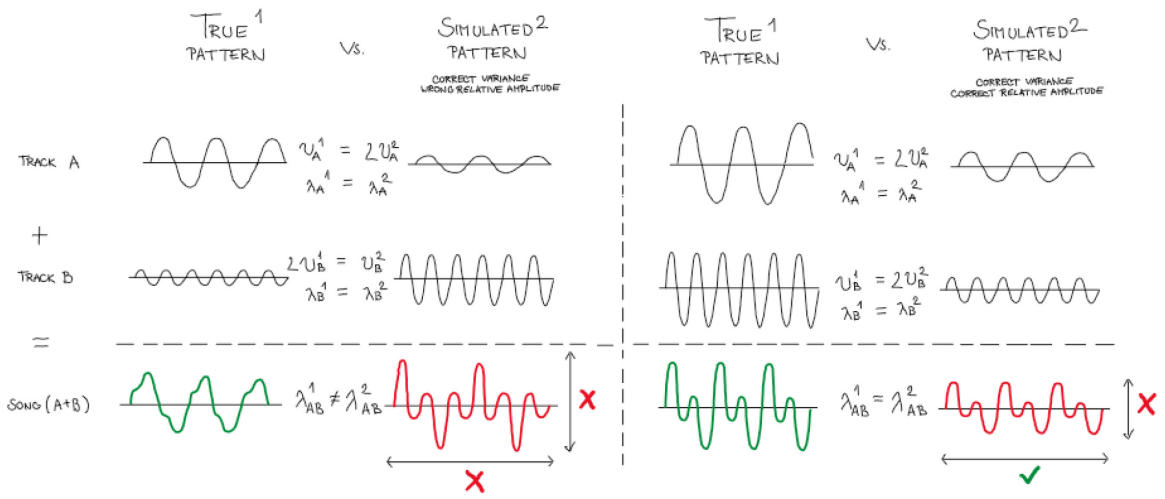


Figure 42. The effect of combining patterns without correctly scaling their amplitude. Left: Relative proportions of the amplitude between the individual patterns A and B are unknown. The combined pattern has the wrong variance and amplitude. Right: Relative proportions of the amplitude between the individual patterns A and B are known. The combined pattern has the correct variance and the wrong amplitude.

To scale the endogenous movement components for different activity types in proper relation to each other, we use the MiD2017 travel data capturing the number of trips attracted by each travel activity type for the entire study area (Appendix 11). In specific, we derive from the MiD2017 study the total number of trips to each of the six travel activities per day in the whole study area (Figure 43). From the total trip count, we calculate the distribution of the individual activity categories (e.g., the relationship of work to education to shopping trips). After calculating the actual trip distribution per activity type, we transform the simulated movement accordingly.

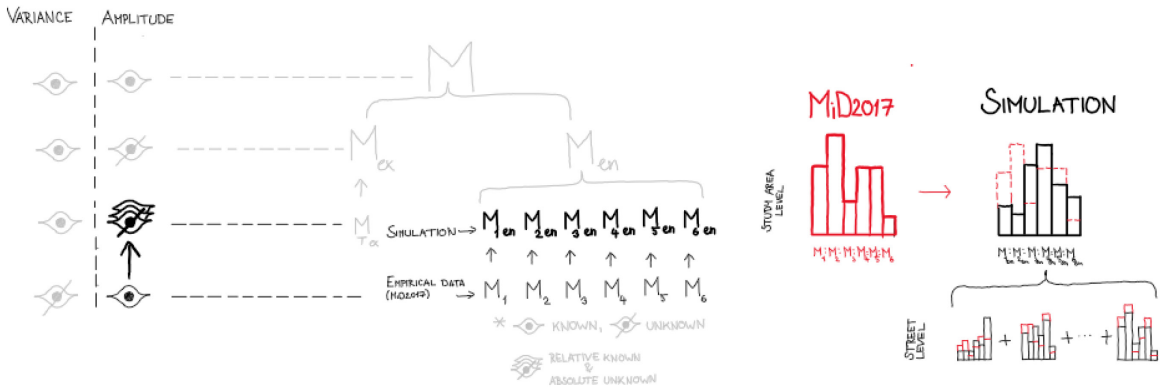


Figure 43. Scaling the amplitude of the simulated endogenous movement components for each activity M_{en}^{T} by MiD2017 movement data. After the scaling procedure, the individual movement components stand in the right relations to each other. However, their absolute amplitude remains unknown.

After this transformation, the amplitude of the simulated movement attracted by different activity types is in proper relation to each other. It means that we can tell if people living

at a specific address walk more often to work or to school. On the other hand, the amplitude of each movement type is still unknown in absolute terms. In other words, we know if people are walking more to work than to school by not how often.

This relative amplitude suffices to derive the variance of the aggregated endogenous movement M_{en}. By simply adding up the transformed patterns of the endogenous movement components $M_{T(en)}$ to get the variance of the aggregated endogenous movement (Figure 44). In other words, we combine the pedestrian movement attracted by all six activity types to get the overall movement pattern generated by activities.

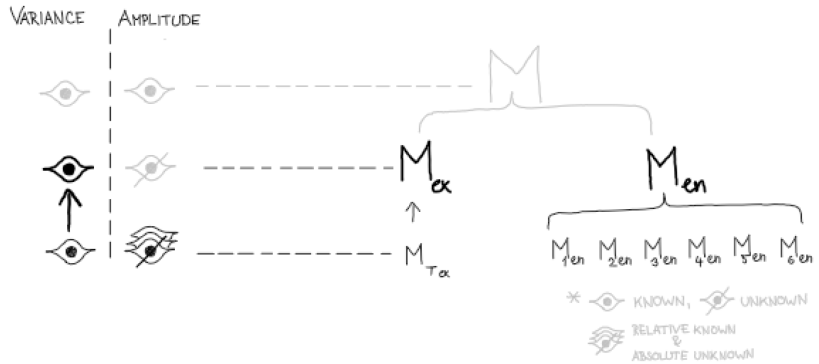


Figure 44. Combining the endogenous movement components by activity type $M_{T(en)}$ to derive the variance of the aggregated endogenous movement M_{en}. Variance in M_{ex} equals the variance in $M_{T(ex)}$.

As next, we estimate the amplitude of the M_{en} and M_{ex} movement pattern by fitting regression equation with the M as the dependent variable and M_{en} and M_{ex} as predictors (Figure 45). Since the total movement M is the only variable with information on the variance and amplitude, we can use it to find the scaling parameters for the rest of the movement component pyramid (Appendix 11). To guarantee that the scaling parameters (i.e., regression coefficients) found by the regression model are only positive (negative movement is conceptual nonsense), we adopt a special version of linear regression – the penalized regression (Appendix 16).

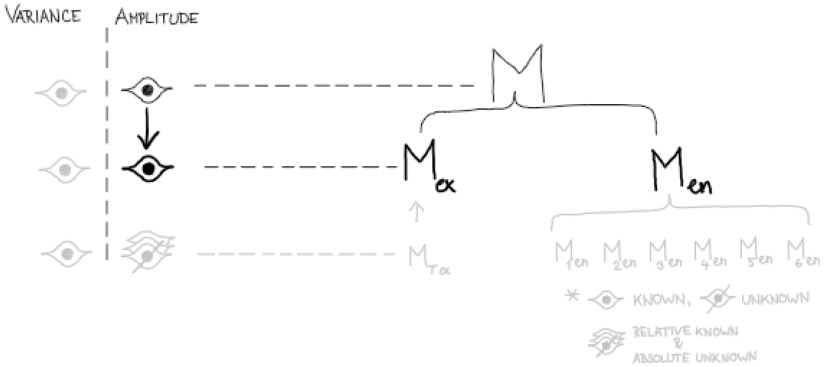


Figure 45. Estimating the amplitude of the M_{en} and M_{ex} from M via the penalized regression.

Finally, we use the estimated scaling coefficient for M_{en} to find the amplitude of the $M_{T(en)}$ components (Figure 46). By doing this, we gain the information on variance and amplitude of the complete movement variable pyramid. Consequently, we are able to compare spatial movement patterns generated by a) different activity types and b) the exogenous or endogenous processes. In other words, we can quantify in absolute terms the pedestrian movement directly caused by the urban form and the movement caused by the allocation of activities.

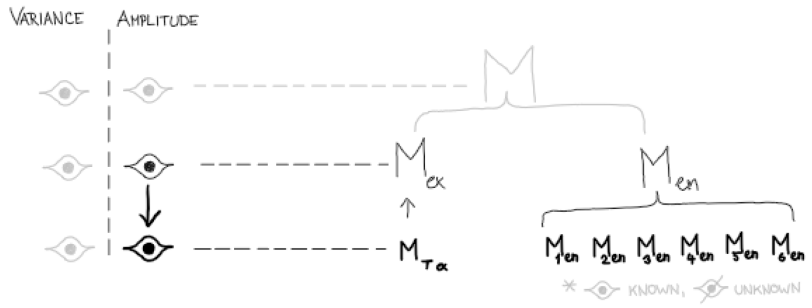


Figure 46. Scaling the relative amplitude of the $M_{T(en)}$ movement components to the absolute movement (person/day).

4.2.5 Common Analysis Unit – Data Aggregation

To model the interactions between movement and activities, we need to map both on the same spatial unit of analysis. To do this, we can a) map street network data on buildings, b) map activity data on the street network or c) map the street and activity data on something else (e.g., neighborhood, analysis grid). Each option represents a specific trade-off between accuracy and simplicity of the resulting model. From a computational point of view, the critical part of the FAMI model is the movement simulation in general and the graph-based calculation of shortest paths in specific. The computational complexity of this calculation follows geometric function (Cormen et al., 2001), which means that the number of calculations is growing much faster than the number of elements in the graph. The current implementation of the DecodignSpaces analysis toolbox for Rhino/Grasshopper used for the shortest path calculations (Appendix 15) is currently limited to roughly 10 000 graph nodes (i.e., street segments). As a result, we use the 7104 street segment as the analysis unit offering the highest possible resolution and acceptable computational complexity and map the 34 871 building objects on it.

We must mention that mapping activities from buildings onto the streets not only reduces the resolution of the activity data but, in some cases, also introduces bias. The problems arise for activities that can be accessed from multiple streets at the same time. Even though several options on how to achieve this are available, they are all problematic in some way. We empirically tested three options (Figure 47) where the problematic activity accessible from several street segments is assigned a) to the closest street, b) to all accessible streets,

or c) divided between all accessible streets. We compared each option, and its resulting bias with the more complex disaggregated model (i.e., each building is a node in the graph) and found that the differences are in the case of Weimar are negligible (Appendix 9). Consequently, for the purpose of this study, we choose the computationally simplest aggregation method – assignment to the closest street.

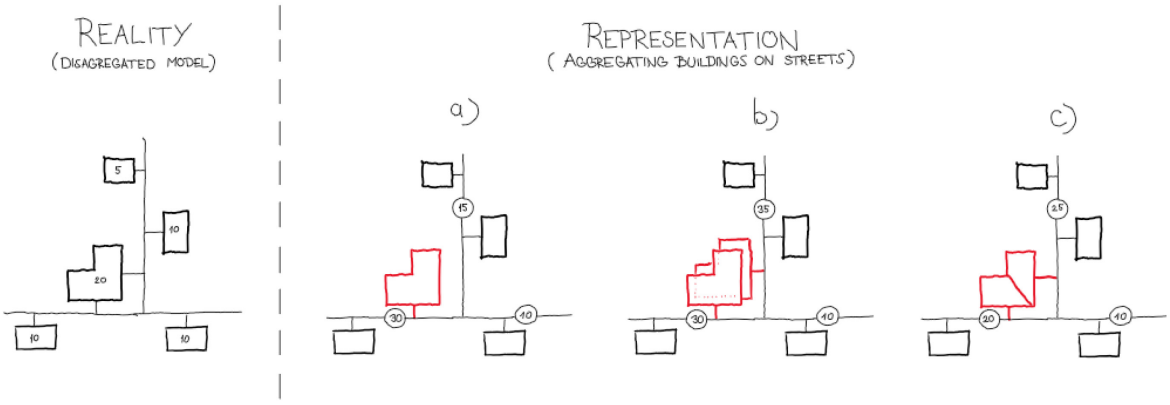


Figure 47. Three different methods for aggregating buildings on street segments. a) closest street aggregation, b) accessibility-based aggregation, c) normalized accessibility-based aggregation.

4.3 Hypothesis Testing

In the following, we discuss the methodological approach for testing the hypothesis H1 and H2 about the significance of individual interactions between FAMI and the validity of models estimating these interactions in isolation.

4.3.1 Testing H1
Movement as a Product of Allocation of Activities and Urban Form

We start with testing the hypothesis H1a, H1b, about the direct effect of urban form and the activity allocation on pedestrian movement. If we find both effects significant, we test H1c for the presence of the omitted variable bias when the effects are estimated in isolation. By doing so, we quantify the validity of the movement model based only on the information on the urban form (i.e., exogenous movement) as used by CUM scholars and planners in the early design stages.

The H1 hypothesis test consists of a two-step sequence. First, we estimate the aggregated exogenous (i.e., being derived from the urban form) and endogenous (i.e., being derived from allocation of activities) movement pattern (Figure 48a). As second, we estimate the contribution of each movement component to the total movement (Figure 48b).

a) b)

Figure 48. Testing the research hypothesis H1a, H1b, H1c, and H1d. a) Estimating the variance in the aggregated endogenous and exogenous movement pattern. b) Estimating the contribution of urban form and allocation of activities to the overall movement.

To quantify the contribution of each movement component to the overall movement, we estimate a series of regression models with M_{ex} and M_{en} as explanatory variables and M as a dependent variable[15].

$$combined\ model: M_{total} = \alpha_{ex}M_{ex} + \alpha_{en}M_{en} + \varepsilon_1 \quad (12)$$

$$exogenous\ model: M_{total} = \alpha_{ex}M_{ex} + \varepsilon_2\ ;\ \varepsilon_2 = \varepsilon_1 + \alpha_{en}M_{en} \quad (13)$$

$$endogenous\ model: m_{total} = \alpha_{en}M_{en} + \varepsilon_3\ ;\ \varepsilon_3 = \varepsilon_1 + \alpha_{ex}M_{ex} \quad (14)$$

We consider the combined model considering all variables simultaneously as the ground truth. We refuse the H1a and H1b if the α_{ex} or α_{en} respectively turn out as not significant. In such a case, we can say that the empirical data does not provide enough evidence to support the hypothesis. If we find the regression coefficients α_{ex} and α_{en} significant, we further investigate their individual contribution to the total movement m. In other words, we measure the portion of total pedestrian movement caused by the allocation of activities and as a direct effect of urban form.

To test for the omitted variable bias hypothesis H1c, we compare the combined model to the exogenous and endogenous model, each ignoring one explanatory variable. If the coefficients α_{ex} and α_{en} are significantly different between the individual and the combined model, we consider the individual model as being biased.

Finally, we test hypothesis H1d, expecting the exogenous and endogenous movement patterns to be significantly different. Here we test for the difference in variance as well as in the amplitude. For this purpose, we run the Pearson's Chi-squared test and compare the spatial distribution of each movement component.

4.3.2 Testing H2
Activity Allocation as a Product of Autocorrelation and Exogenous and Endogenous Movement

In the following, we discuss the methodology for testing the hypothesis H2a, H2b, and H2c about the effect of autocorrelation and the exogenous and endogenous movement components on the allocation of activities (Figure 49). From here, we move to test the H2d – omitted variable bias introduced by estimating these effects in isolation and not simultaneously. Since we expect the activity type to influence these effects, all hypotheses under H2 are tested individually for each of the six relevant travel activities.

[15] We have to note that the regression model used in the hypothesis H1 testing and the regression model used to derive the amplitude of M_{ex} and M_{en} discussed in previous section is the same penalized regression model (see Appendix 11) and thus, both goals can be achieved in single step.

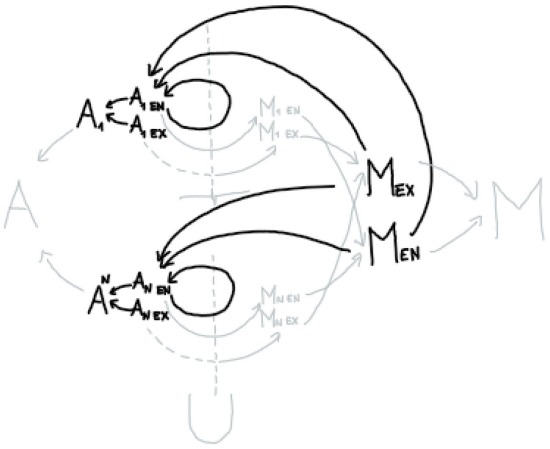


Figure 49. Testing the hypothesis H2a, H2b, H2c, and H2d.

From the conceptual standpoint, we adopt a similar approach to the testing of H1. We estimate the regression model with exogenous activities $a_{T(ex)}$ and endogenous activities $a_{T(en)}$ as explanatory variables and the overall activity pattern a_T as a dependent variable[16].

$$a_T = \beta_{T(en)} a_{T(en)} + \beta_{T(ex)} a_{T(ex)} + \mu_1 \tag{15}$$

To introduce the effect of movement and autoregression on endogenous activities, we substitute the $a_{T(en)}$ by the following term:

$$a_{T(en)} = \rho W a_{T(en)} + \alpha_{T(ex)} m_{ex} + \alpha_{T(en)} m_{en} + \mu_2 \tag{16}$$

The resulting regression equation to estimate is accounting for the spatial autoregressive term $\rho W a_{T(en)}$, endogenous movement $\alpha_{en} m_{en}$, exogenous movement m_{en}, and exogenous activities $a_{T(ex)}$.

$$a_T = \beta_{T(en)} (\rho W a_{T(en)} + \alpha_{T(ex)} m_{ex} + \alpha_{T(en)} m_{en} + \mu_2) + \beta_{T(ex)} a_{T(ex)} + \mu_1 \tag{17}$$

Since the regression model aim is to explain the variance, we can ignore the exogenous activity component $a_{T(ex)}$ as it is constant (i.e., its variance is zero). Additionally, we have to keep in mind that a_T and $a_{T(en)}$ are identical in terms of their variance, which leads us to further simplification (see Chapter 4.2.2.1). We substitute the previously introduced regression model for testing H2 with a more straightforward but equivalent model. It

[16] In case of spatial autoregressive model, we adopt a matrix notation. This means that vector variables are marked with lowercase letters and matrix variables with upper case letters. The m_{ex} and m_{en} variables in matrix notation represent the same variables were previously marked as M_{ex} M_{en}.

contains m_{ex}, m_{en} and autoregressive term $Wa_{T(en)}$ as explanatory variables and $a_{T(en)}$ as the dependent variable (Figure 50).

$$combined\ model: a_T = \rho W a_{T(en)} + \alpha_{T(ex)} m_{ex} + \alpha_{T(en)} m_{en} + \mu \quad (18)$$

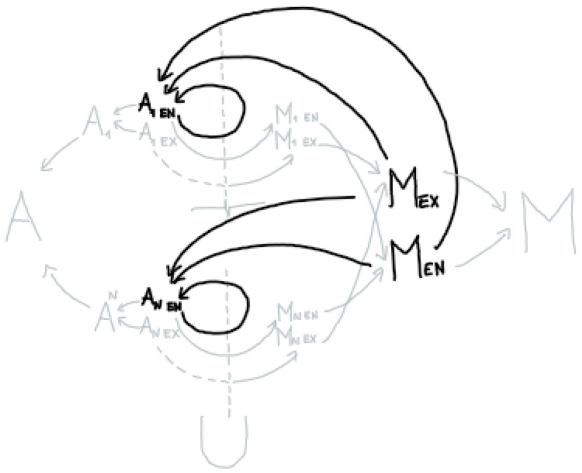


Figure 50. Simplified testing framework the hypothesis H2a, H2b, and H2c.

From here on, we construct a series of sub-models using a reduced set of explanatory variables. These are used to compare the estimated coefficients $\alpha_{T(ex)}$, $\alpha_{T(en)}$ and quantify the omitted variable bias.

$$sub-model\ 1: a_T = \rho W a_{T(en)} + \alpha_{T(en)} m_{en} + \mu \quad (19)$$

$$sub-model\ 2: a_T = \rho W a_{T(en)} + \alpha_{T(ex)} m_{ex} + \mu \quad (20)$$

$$sub-model\ 3: a_T = \alpha_{T(ex)} m_{ex} + \alpha_{T(en)} m_{en} + \mu \quad (21)$$

$$sub-model\ 4: a_T = \alpha_{T(en)} m_{en} + \mu \quad (22)$$

$$sub-model\ 5: a_T = \alpha_{T(ex)} m_{ex} + \mu \quad (23)$$

4.3.2.1 Statistical Model

The statistical model used to estimate the allocation of activities (H2) differs from the model used to estimate the distribution of movement (H1) in two major ways. The first specialty is that it accounts for spatial autoregression between activities themselves. This relationship suggested by urban economists is particularly challenging to model as the interaction between activities brings the same variable on both sides of the equation. The difficulty lies in the attempt to model the interaction between activities without having interaction data[17].

[17] All collected data is cross-sectional. In other words, they capture a single moment in time. Interactions are however a temporal process which is best represented by longitudinal data.

We discuss this problem and the approach of spatial econometrics to deal with it in detail in Appendix 18, but in essence, the solution is to introduce the spatial weights matrix W to simplify the estimation of the interaction matrix by imposing a structure of spatial relationships.

The second distinctive feature of the statistical model used in H2 is related to the activity data distribution. In the case of activities, we observe a binomial distribution, with the main peak at zero activity intensity followed by a normal distribution (Figure 51).

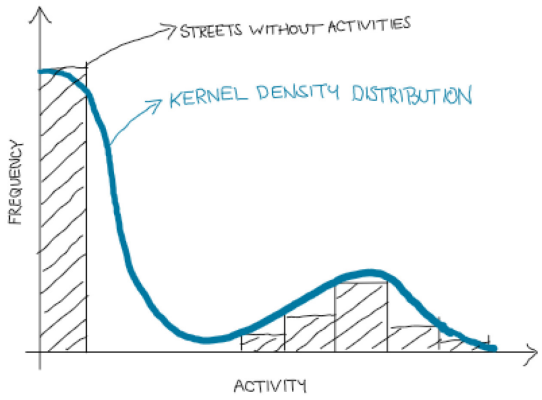


Figure 51. Illustration of bimodal density distribution of activity intensity.

In general, multimodality suggests that multiple processes are taking place at the same time. In the case of activity allocation, we assume that a) the change in the pedestrian movement is linearly related to change in activity allocation intensity; however, b) some minimal threshold level in movement is required. To illustrate this point, we expect the number of restaurants per street to grow linearly with increasing movement flow. However, first, some critical number of passing-by pedestrians must be reached. As a result, the statistical model used to test the H2 hypothesis is based on two sub-models targeting the two processes:

a) *Filter* - finding the critical threshold in movement required for the emergence of activities. We estimate this threshold via logistic regression and filter-out the street segments with zero activity levels.
b) *Amplifier* - modeling the linear relationship between movement and activity for street segments, which are above the critical threshold value determined in the *Filter*. The relationship is estimated via a linear spatial autoregressive model.

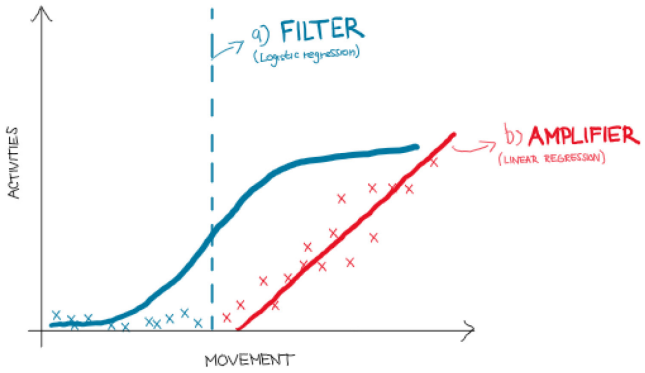


Figure 52. Illustration of the two-step statistical model consisting of a) Filter (logistic regression) and Amplifier (linear spatial autoregressive model) component.

4.3.2.2 Variable Normalization

To estimate the effect of autocorrelation and movement on activity distribution, we must correct for the variation in the street segment capacity. In other words, if two streets offer the same potential (e.g., pedestrian flow) for establishing an activity, naturally, if one street is longer than the other, we will expect more activities at the longer street. To guarantee that what we are modeling is the effect of movement and not just a difference in the street capacity, we must take it into account.

Such normalization can be done either by a) dividing the activity intensity or b) multiplying movement flows (i.e., *Through-Frequency*) by street capacity [18]. The former approach results in a point measure, while the later adopt the line as a geometrical unit of measurement (Figure 53). Both options result in equivalent prediction models differing only in their measurement units. For practical purposes of model interpretation, we employ the point approach to the *Filter* model and the linear approach to the *Amplifier*.

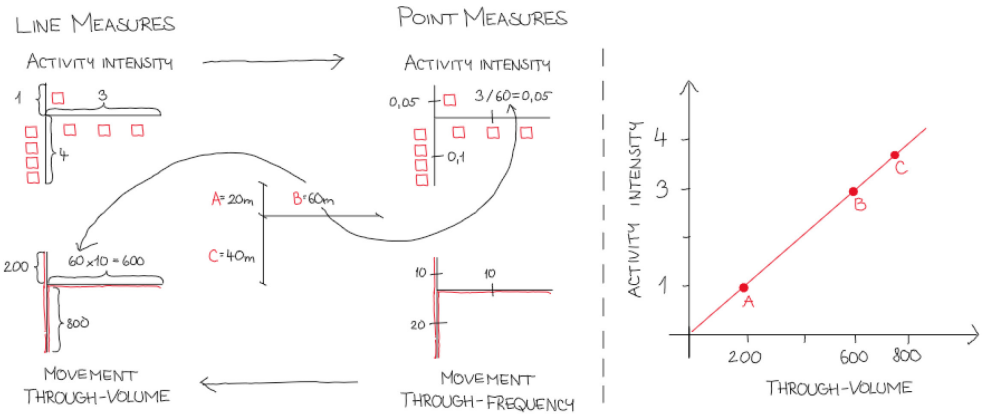


Figure 53. Activity and movement normalization. The transformation from line to point measures and back.

[18] We approximate the street capacity to accommodate activities as function of its length.

Finally, we address the skewed and non-normal distribution of both explanatory (i.e., exogenous and endogenous movement) and dependent variables (i.e., activity intensity). To improve the overall accuracy and reliability of the linear regression model, we logarithmically transform all variables so they follow normal, or close to normal distribution. As suggested by O'Hara & Kotze (2010), we adopt logarithmic transformation $\log(x + 1)$ to reduce the bias in the regression coefficient estimation and improve the model performance.

4.4 Limitations

In the following, we discuss the impact of data availability and methodological limitations on our ability to test the research hypothesis. We identify the following restrictions constraining the accuracy of the interaction model.

a) Not all travel activities are considered. The six travel activities represented in the interaction model account for 45% of all trips in the study area (Appendix 7).
b) The movement model represents only single-purpose trips, no trip chains. In the case of pedestrian movement, Primerano and colleagues (2008) found that 88% of pedestrian journeys are single purpose.
c) The movement model does not consider the impact of topography. We note that the study area is composed of predominantly flat terrain, with 95% of travel activities location between 220m and 250m elevation.
d) Building typology is not considered. The interaction model is treating every building as being able to accommodate any of the selected six activity types.
e) Street capacity to allocate activities is not considered. In other words, we do not consider the competition between activities for limited space. We found that the mean utilization per street segment is 22.9%, which means that, on average, 77.1% of the ground floor area was not used by any travel activity other than accommodation.
f) The movement model is assuming prior knowledge of the environment. Thus it is not capable of capturing movement patterns of tourists or other pedestrians unfamiliar with the local urban context.
g) The movement model is based on cognitive shortest paths only. This covers 85% of the movement in the empirical study conducted by authors (Appendix 13).
h) The FAMI model is simultaneous (i.e., activity affects movement and movement affects activity). However, the regression models used to test the H1 and H2 do not correct for simultaneity bias. As discussed in Appendix 3, the simultaneous relationships are non-linear, and thus it does not threaten the validity of the linear models employed in the hypothesis testing.

We argue that each of the listed methodological limitations might have a significant impact on the predictive power and accuracy of the FAMI model. Nevertheless, we do not observe any of the limitations to cause a systematic error, which would challenge our ability to answer the research questions and test the research hypotheses H1 and H2. We emphasize that this study is not focused on the prediction accuracy but testing the overall validity of current models explaining the form-activity-movement interactions. Thus, we conclude that despite the above-mentioned limitations, this study's overall aim has not been compromised.

5 Results

5.1 Movement - Testing Research Hypothesis H1

In the following, we test the research hypothesis H1 about the direct effect of urban form and allocation of activities on pedestrian movement. Furthermore, we test if both effects must be estimated simultaneously in order to prevent the omitted variable bias.

We start by estimating the variance and amplitude of all components of the movement variable pyramid (Figure 32). We accompany each step with a schematic diagram referencing the respective section of the "Research Methods and Data" chapter.

5.1.1 Variance and Amplitude of Movement Components

Empirical Data on Total Movement

We start by discussing the empirical data on movement *Through-Frequency* and present the results of the temporal transformation procedure applied to the data. The empirical pedestrian counts (see Appendix 7) are transformed from the original 15 minutes counting interval to pedestrian frequency per day by considering the variation in movement flows during the daytime (Figure 54). By doing so, we improve the interpretability and representativeness of the study results by keeping the data acquisition effort relatively low (i.e., we get estimates of 24 -hour pedestrian counts by spending only 15 minutes at each location).

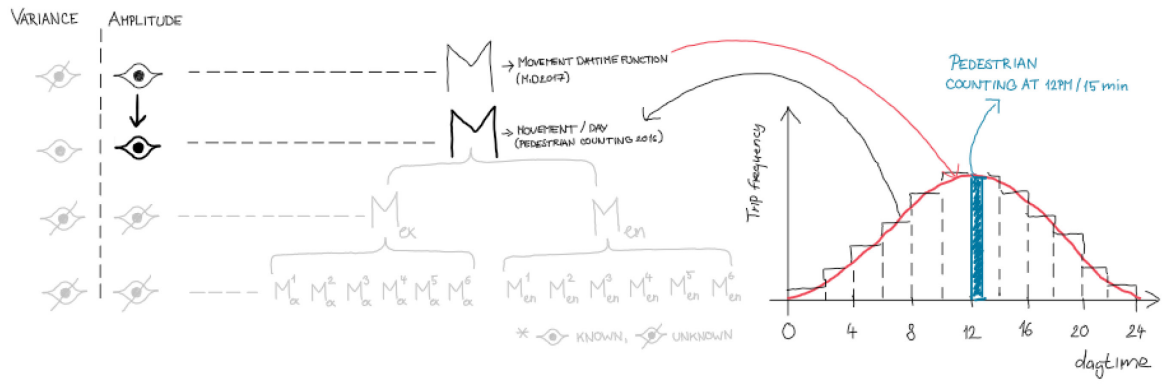


Figure 54. Scaling the amplitude of the pedestrian movement pattern (Through-Frequency) from the original 15 minutes counting interval to the whole day.

As described in Appendix 7, we measured pedestrian counts at 100 locations across the study area. The counts are ranging from 1 up to 293 pedestrians per 15 minutes period[19]. We observe the high pedestrian frequency (above 100 pedestrian/15min) to be concentrated around the historical city center and train station. The residential neighborhoods have significantly lower pedestrian frequencies ranging from 1 to 20 pedestrian/15min for most locations. The single exception is the large condominium estate in the north-west of the study area (Figure 55). Here we found higher pedestrian frequencies going up to 40 pedestrian/15min.

Figure 55. Distribution of empirical pedestrian counts (Pedestrian counting, 2016). The size of the circle represents the average number of passing-through pedestrians per 15-minute counting time.

[19] Pedestrian counts at each location were repeatedly measured nine times, the presented results are taken from the average of the nine measurements.

To transform the 15-minute counts into daily pedestrian frequencies, we derive the daily pedestrian movement frequency function for Weimar from the MiD2017 travel data (N = 248). In essence, we calculate the distribution of pedestrian trips throughout the day and calculate the respective density function. The resulting histogram and density plot (Figure 56) confirm our previous findings and show that the amount of traffic changes significantly over the daytime. We see a steep rise in movement starting at 6 am and peaking at 11 am. This peak is followed by a short drop in the movement at 12 am and another rise peaking at its daily maximum at 5 pm. After this second peak, the pedestrian movement constantly drops toward zero at 2 am.

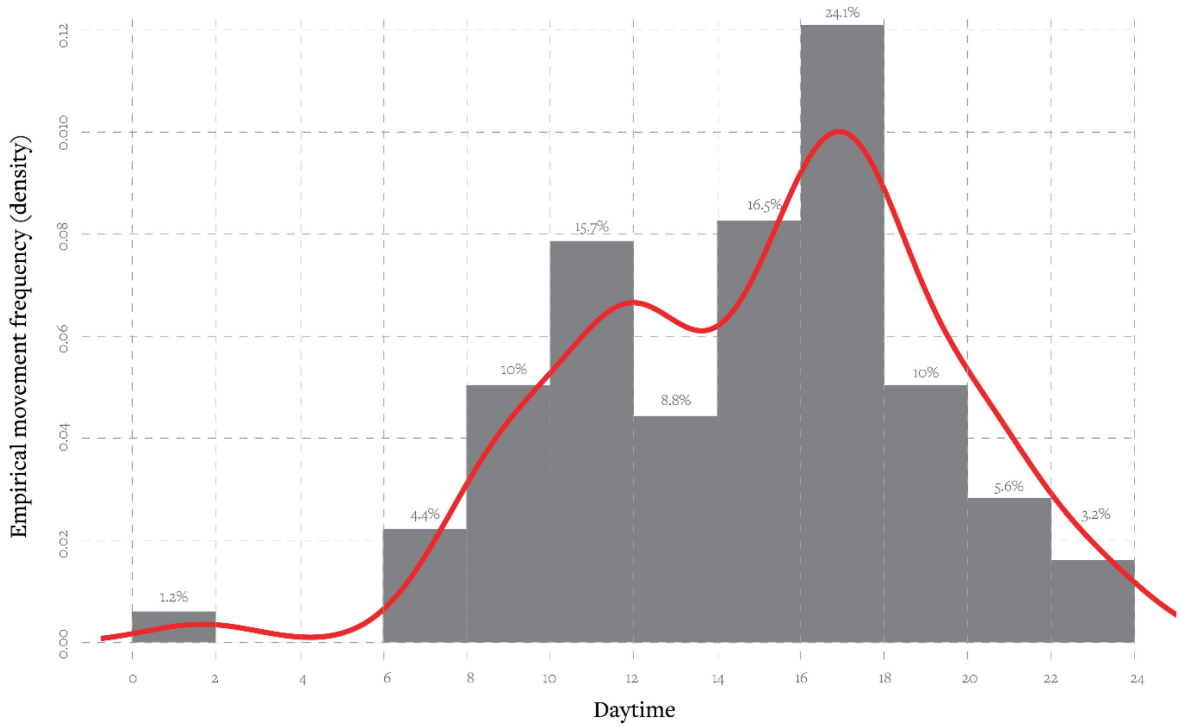


Figure 56. Histogram and density plot of pedestrian movement frequency in Weimar across daytime as captured by the MiD2017 study. 248 pedestrian paths in Weimar from MiD2017 study are clustered in 12x2 hours bins (e.g., 0-2 am, 2-4 am, etc.) to match the counting windows of our pedestrian frequency study (8-10 am, 12-2 pm, 4-6 pm). The red line shows the kernel density estimates [20]

Based on the density function derived from the pedestrian movement distribution in the MiD2017 data set (see red line in Figure 56), we estimate the hourly and daily pedestrian frequency for all 100 counting locations (Figure 57). We found that 1370 pedestrians pass through the selected street segments on average, with the minimum and the maximum number of pedestrians being 54 and 14944, respectively.

[20] Default R (R Core Team, 2017) function density is applied with Gaussian kernel and bandwidth equal to standard deviation of the kernel.

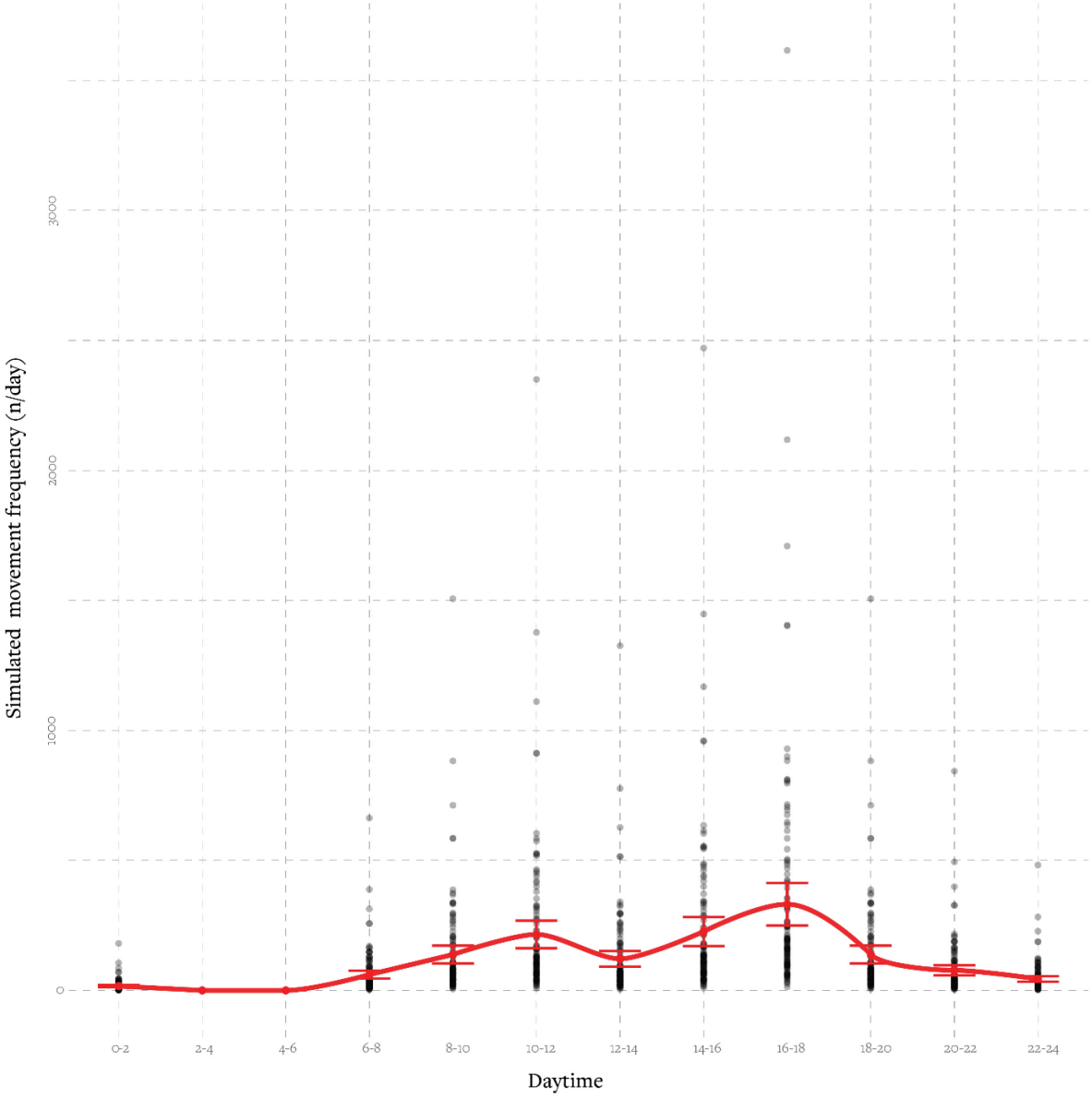


Figure 57. Estimated hourly distribution for the pedestrian movement frequency for 100 segments observed in the empirical study. The estimated frequencies are binned in 12 x 2 hours bins. The grey points are the estimates for individual streets for each time bin, and the red dot shows the mean frequency with its 95% confidence interval.

Simulated Endogenous and Exogenous Movement by Activity Type

The direct effect of urban form and allocation of activities on pedestrian movement was derived computationally via movement simulation model (Figure 58). The movement model was based on a street segment map weighted by the six travel activity types at the destination and the accommodation at the origin of the movement. The path between each

origin-destination pair is minimizing the cognitive distance (i.e., angular deviation) (Appendix 13). The travel impedance function utilizes the gravity accessibility with a negative exponential function and the empirically calibrated beta coefficient $\beta = 0.00138$ (Appendix 14).

By applying different destination weights to each travel activity type, we end up having six distinct models for the endogenous pedestrian movement. Additionally, we estimate one exogenous pedestrian movement model.

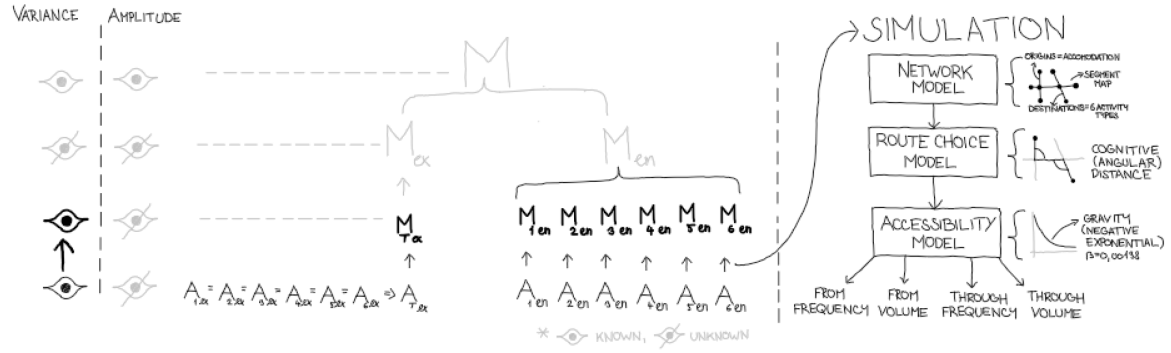


Figure 58. Simulating variance in the exogenous and endogenous movement pattern by activity type.

For each movement model, we calculate the four movement characteristics (i.e., *Through-Frequency, Through-Volume, From-Frequency, From-Volume*), capturing different aspects of how pedestrians move across the study area (Figure 59). Even though the in-depth analysis of all 28 movement patterns is beyond the scope of this study, we can confirm significant variation in the movement patterns within as well as between the activity types and movement characteristics. These findings are in line with the results of the exploratory factor analysis, suggesting that each of the six activity types should be treated as a distinct variable. Moreover, it supports the argument that movement is a multidimensional concept with multiple characteristics, each describing a different aspect of how and why people move.

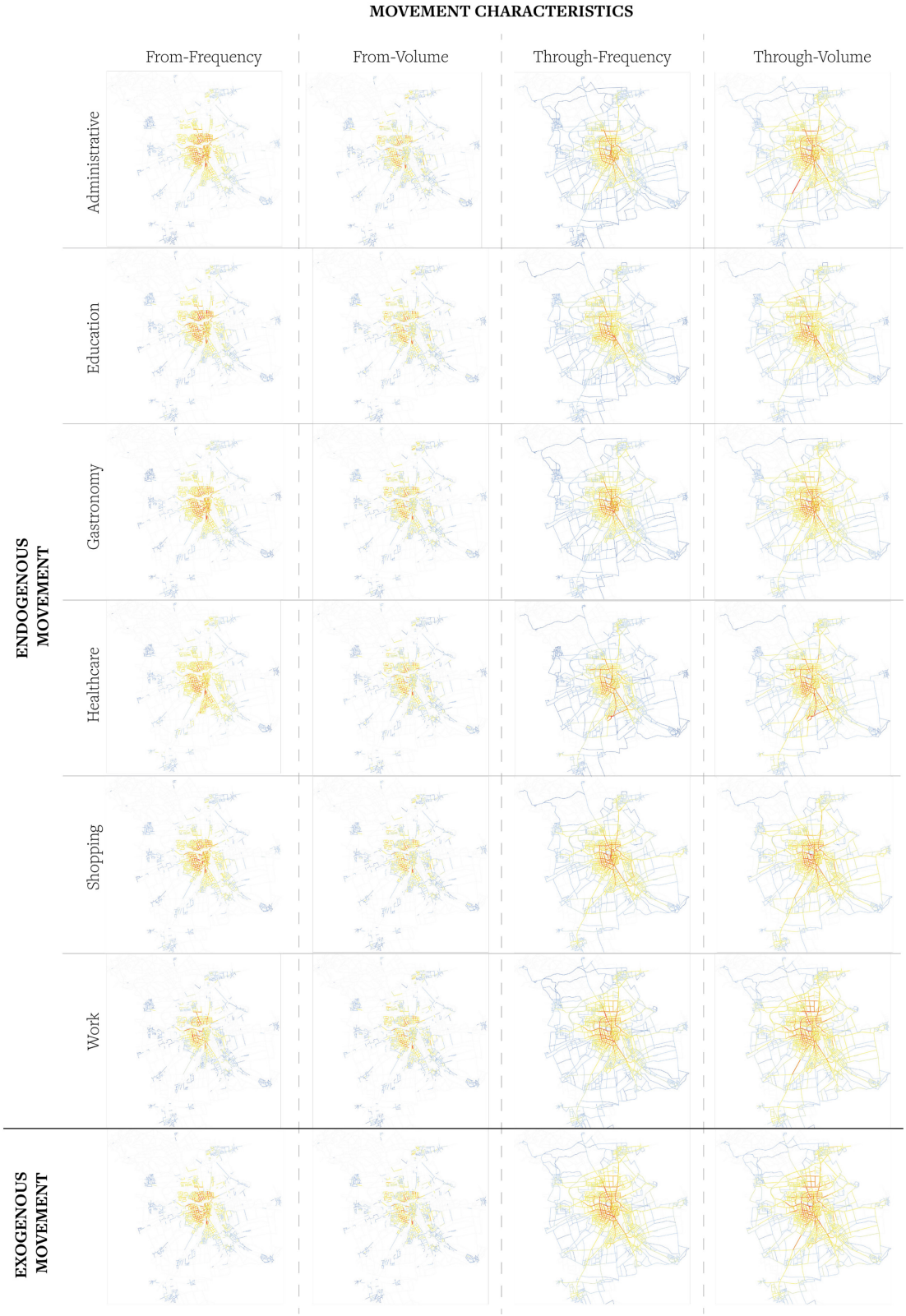


Figure 59. Simulated exogenous and endogenous movement per activity. Each movement pattern is represented by the four movement characteristics.

After simulating the pedestrian movement attracted by each of the six activities, we aggregate the individual movement patterns into the total endogenous movement. The aggregated endogenous movement depicts the overall movement caused by the allocation of activities, and together with the movement directly derived from urban form (i.e., exogenous movement) represents the basis for testing the research hypotheses H1a, H1b, H1c, and H1d.

In order to aggregate the movement attracted by individual activities, we must normalize their amplitude. The reason why we cannot simply combine all six patterns is that the destination weights in the movement simulation model are based on their relative and not absolute attractiveness. In other words, we assume that the more a destination has to offer, the more movement it attracts, but how much remains unknown. The proxy of this relative attractiveness of any activity is its floor area (e.g., we expect a larger store to attract more shoppers) [21].

The difficulty arises when comparing different activities as we expect the same floor area of shopping and administrative activities to attract a different amount of movement. The argument is perhaps most obvious when looking at spaces where multiple activities take place at the same time. For instance, in most shops, the difference between the number of employees (i.e., workers) and customers (i.e., shoppers) can be counted in orders of magnitude. However, if we approximate the amount of work seeking pedestrians and shoppers based on the same floor area, we ignore this difference. To account for the varying attractivity of different activities, we normalize the relative activity weightings so that each movement pattern follows the empirical distribution of trips by activity in the study area (Figure 60). The empirical distribution has been derived from the travel diaries collected in the scope of the MiD2017 study and represents the total portion of pedestrian trips by activity.

[21] Even though the activities of the same type with the same floor area might marginally differ from each other (e.g. not all bakeries of the same size are equally attractive), we consider this as acceptable approximation.

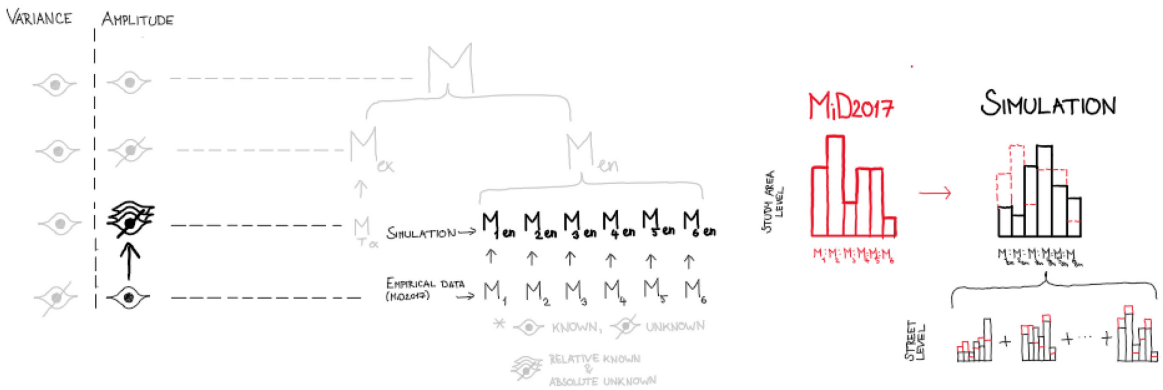


Figure 60. Scaling the amplitude of the simulated endogenous movement components for each activity M_{en}^{T} by MiD2017 movement data. After the scaling procedure, the individual movement components are in the correct relationships to each other. However, their absolute amplitude remains unknown.

We found that in Weimar, the six activity types considered in this study account for 45% of all pedestrian trips (Figure 61). By far, the largest portion of trips is attracted by shopping activity (16.9%), followed by education (8%), work (6%), health (5.8%) and gastronomy (3.9%).

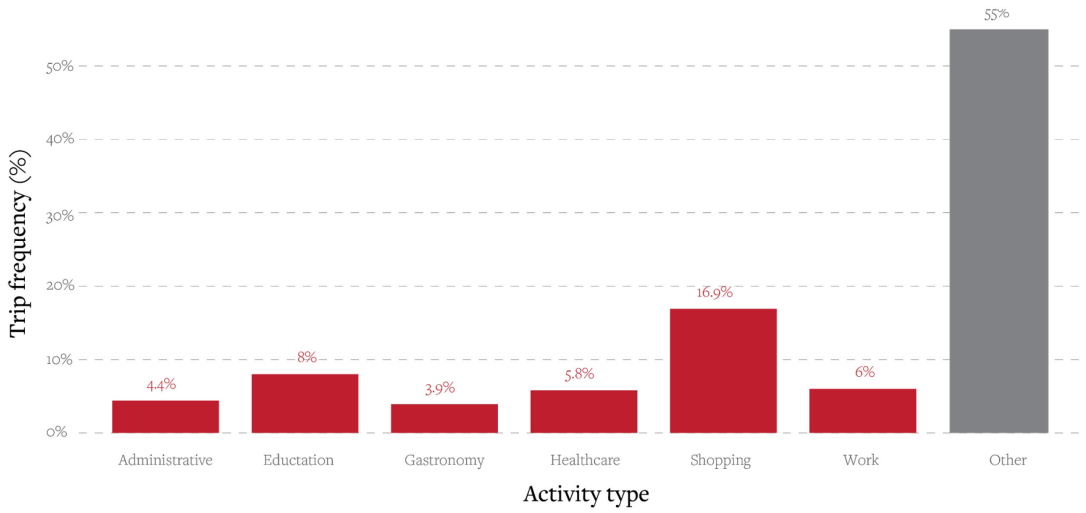


Figure 61. Distribution of trip frequency per activity (MiD2017 study).

After finding the scaling coefficient for the attractivity weighting of each movement activity type, we apply it to all four movement characteristics. In the following, we explore the effect of activity type on the movement volume and movement frequency. For this purpose, we compare the distribution of the total *From-Frequency, From-Volume,* and *Through-Volume* per

activity[22]. As discussed in Appendix 10, the *From-Volume* and *Through-Volume* are equivalent in their sum; thus, we expect both to follow the same distribution. Pearson's Chi-square test shows that both distributions are not significantly different ($\chi^2 = 4.441$, df = 5, p-value = .49) and confirm the theoretical assumptions. The minor differences between total trip and traffic volume per activity (Figure 62) can be attributed to artifacts in the implementation of the movement simulation engine (see Appendix 16).

From here, we move to a comparison of the difference between the *From-Frequency* (i.e., number of trips per day) and *From-Volume* (i.e., distance of trips per day) distribution. If the average trip distance for each activity would be the same, we expect these two distributions to be equal. However, as shown in Figure 62, the distributions are significantly different ($\chi^2 = 133.64$, df = 5, p-value = $2.2e^{-16}$). We observe that the work and education activities are producing a higher portion of trips than the traveled distance. In other words, trips to these activities are shorter than the average journey in Weimar. On the contrary, the pedestrian movement to shopping, administrative, gastronomy, and healthcare activities show the opposite trend.

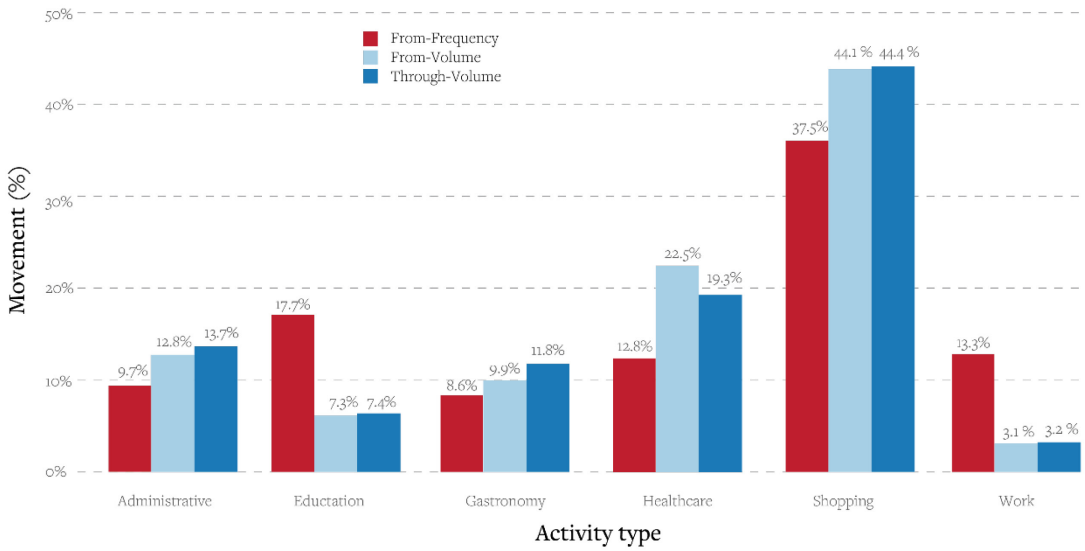


Figure 62. Comparison of total endogenous Through-Volume, From–Volume and From-Frequency per activity.

[22] We do not consider the Through-Frequency as its sum depends on the arbitrary decision on the number of counting locations. It is similarly meaningless as measuring wind speed at different locations and than adding it together. On contrary, the From-Volume, From-Frequency and Through-Volume consider the capacity of each street segment in terms of its length or associated floor area and thus are representative for the system as whole. In essence, if we would arbitrary subdivide the existing street network (splitting long street segments in to several shorter ones) it will have no impact on the total From-Volume, From-Frequency and Through-Volume but the Through-Frequency would change.

These differences between activities can be traced to the varying average distance from home location to possible destinations. More centrally allocated and frequently distributed activities need less travel distance to reach them. Since work activities are most frequent of all activities and distributed throughout the city, they are on average closer than something less frequent or central such as healthcare.

We illustrate these differences by calculating the traffic volume for each activity by holding the trip frequency constant for all of them as depicted in Figure 63. If each activity accounted for the same portion of trips (1:6 =16,6%), the work and education would produce only 4 and 7 percent of traffic volume respectively. The shopping, administrative, healthcare, and gastronomy would produce between 18 and 25 percent.

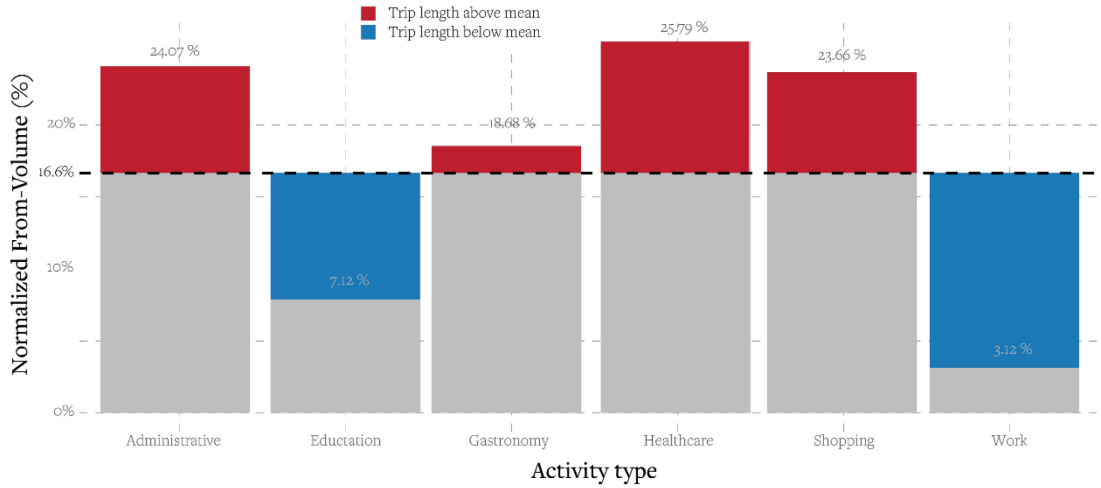


Figure 63. Comparing the trip volume (i.e., total travel distance) per activity while holding the trip frequency (i.e., total numbers of trips) equal.

Aggregated Endogenous and Exogenous Movement

After correcting for the relative differences in the activity weights, we aggregate the simulated endogenous movement components (Figure 64). The aggregation is done by the simple addition of the movement patterns and is conducted for each characteristic (i.e., *From-Frequency, From-Volume, Through-Frequency,* and *Through-Volume*).

The resulting aggregated endogenous movement is described only in terms of its variance. The amplitude (i.e., number of pedestrians or travel distance per day) will be estimated in the scope of the hypothesis testing. The variance of the aggregated exogenous movement equals the exogenous movement components per activity, and thus, it is already known from the simulation.

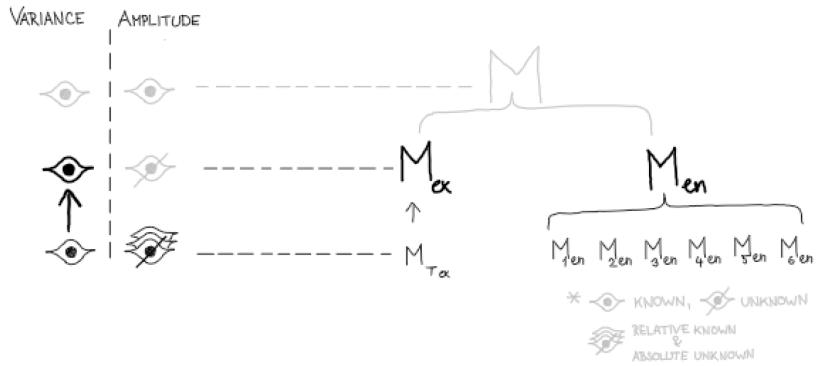


Figure 64. Combining the endogenous movement components by activity type $M_{T(en)}$ to derive the variance of the aggregated endogenous movement M_{en}. Variance in M_{ex} equals the variance in $M_{T(ex)}$.

Hypothesis Testing

By simulating the aggregated endogenous and exogenous movement, we are able to split the variance in the pedestrian movement pattern into two components – the exogenous which can be derived directly from urban form, and the endogenous, which needs additional information on the distribution of activities. The question to be answered is what proportion of the total pedestrian movement can be attributed to each movement component. In other words, we are interested in the amplitude and so the absolute contribution of each movement component to the overall movement. This would allow us to quantify the proportion of the movement, which can be directly assessed from the urban form, and the proportion of movement which needs additional information on activity allocation.

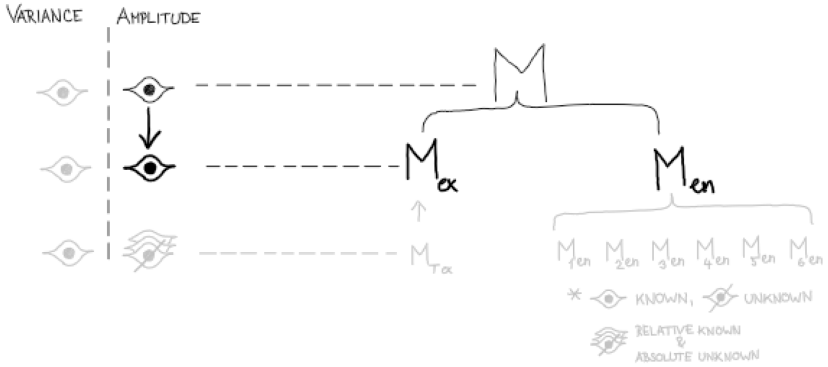


Figure 65. Hypothesis testing and estimation of the amplitude of the M_{en} and M_{ex} from M via the penalized linear regression.

The hypothesis H1a and H1b expect both, the exogenous and endogenous movement components to significantly contribute to the overall pedestrian movement pattern. According to the hypothesis H2c, we assume that both movement components must be estimated simultaneously in order to avoid the omitted variable bias. In other words, we expect that the estimated contribution of each movement component will be systematically

biased and thus wrong if it is estimated alone. Finally, we test in H1d if the exogenous and endogenous movement patterns significantly differ.

For the purpose of hypothesis testing, we estimate the amplitude of the M_{en} and M_{ex} components by fitting the regression model (Equation 24) with the M as dependent and M_{en} and M_{ex} as explanatory variables. We represent the simulated movement components by their *Through-Frequency* characteristic as the empirical data for the total movement pattern M was collected in the form of pedestrian counts (Appendix 7). The estimated regression coefficients α_{ex} and α_{en} are scaling parameters which reveal the missing information about the amplitude of the movement components M_{en} and M_{ex}. In other words, the product of $\alpha_{ex}M_{ex}$ and $\alpha_{en}M_{en}$ reveals not only how movement varies throughout space but also its absolute values[23] (e.g., person or kilometer per day) at each location.

$$M = \alpha_{ex}M_{ex} + \alpha_{en}M_{en} + \varepsilon \tag{24}$$

To guarantee that only positive movement can be estimated, we adopt a special type of linear regression model – the penalized regression. The estimation details are discussed in Appendix 11, but in essence, we restrict the range of each parameter, so only positive movement estimates are allowed to fulfill the conceptual constraints of the model.

In the following, we present the results of the penalized ridge regression explaining the total movement frequency per day measured for 100 street segments as a linear combination of the simulated endogenous and exogenous movement frequency. In penalized regression, the amount of the penalty must be fine-tuned using a lambda Λ constant. When Λ is equal to 0, the penalty term has no effect, and ridge regression will produce the classical least square coefficients (i.e., standard linear regression). However, as Λ increases to infinite, the impact of the penalty grows, and the ridge regression coefficients will get close to zero. We find the optimal value of lambda we minimize the cross-validation error as proposed by (Hastie et al., 2009). Here we look for the lambda value, which fulfills the parameter constraints and minimizes the Mean-squared error of the regression, or in other words, it maximizes the model accuracy, as shown in Figure 66. The optimal lambda was found at $\Lambda = 31.907$.

[23] To simplify the interpretation of the regression coefficients, we adjust the both exogenous and endogenous movement components to follow the same mean.

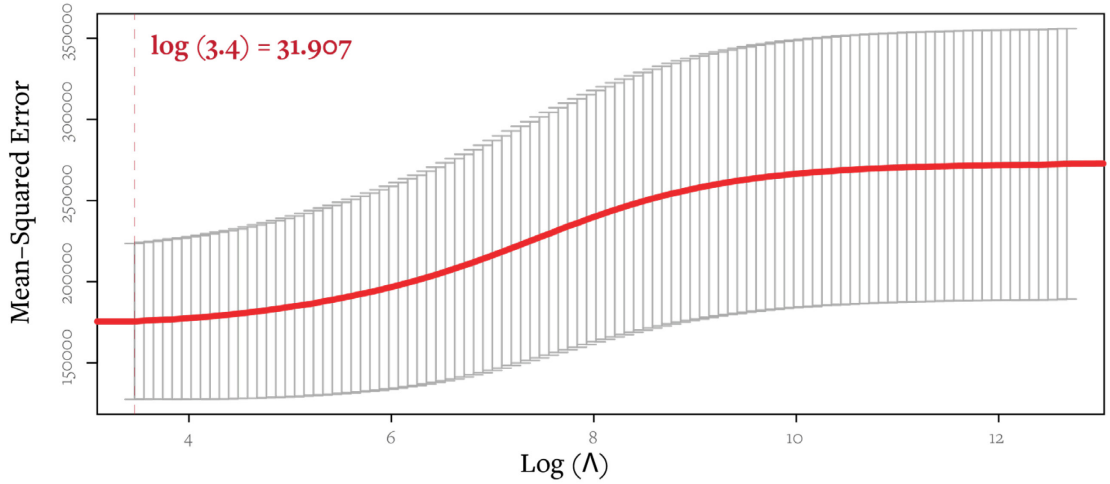


Figure 66. Mean squared error as a function of the lambda coefficient. The left vertical line marks the optimal lambda value.

Hypothesis H1a and H1b

After fitting the penalized regression model, we found the exogenous movement coefficient $\alpha_{ex} = 0.06$ and endogenous $\alpha_{en} = 0.35$ With both coefficients being highly significant ($p < 0.01$), this means that on average, for one pedestrian moving toward an exogenous activity, we expect 5.7 pedestrians are moving towards an endogenous activity[24]. Or in other words, the endogenous movement contributes with 5.7 more pedestrian frequency at the 100 selected segments when compared to the exogenous movement. The results confirm the hypothesis H1a and H1b and can be interpreted as empirical evidence about the significance of both, the exogenous and endogenous movement components. We conclude that they both significantly contribute to the overall movement even though, each by a different amount. Our study suggests that the amount of movement generated directly by the distribution of the urban form is a small fraction of the amount of movement generated by the distribution of activities (~1/6)

Furthermore, we report that the penalized regression explains 58% of the variance (r-squared = 0.58) in the empirical movement as compared with 60% of variance explained by standard linear regression (r-squared = 0.6). It must be noted that the penalization will always reduce the model accuracy as a price for imposing constraints on the model. However, in our case, the drop in the explanatory power is only minor and expresses the amount of bias introduced by the penalty term lambda. Overall, we conclude that there is a large portion of the

[24] The coefficients absolute values have no substantial interpretation as the underlaying exogenous and endogenous movement patterns depict only variation and not the amplitude. However, due to normalization of both variables to the same mean the coefficients can be interpreted in terms of their mutual relationship - fraction ($\frac{\alpha_{en}}{\alpha_{ex}} = \frac{0.06}{0.35} = \frac{5.7}{1}$).

pedestrian movement (42%) not being captured by neither exogenous nor endogenous simulated movement components. We assume that the unexplained variance can be attributed to the fact that the simulated endogenous movement covers only a fraction of all activities (i.e., six activities accounting for 45% of the total number of trips). Although other limitations of the movement-activity interaction model discussed in Chapter 4.4 can play a significant role, we argue that for the purpose of this study, the explained variance is only secondary. The primary goal is to fit and compare the coefficients for the two captured movement components, and this can be accomplished even with a large portion of the variance remaining unexplained.

Hypothesis H1c

To test the hypothesis H1c, we assume that the direct contribution of urban form and allocation of activities on pedestrian movement must be estimated simultaneously. Or, to put it differently, we expect that the individual estimation of each coefficient will lead to omitted variable bias. In general, the bias occurs when we ignore relevant variables that are correlated with some of the explanatory variables included in the model. Without going into technicalities, what happens here is that the effect which comes from the ignored variable is mistakenly attributed to the correlated variable in the model (Appendix 2). If both, the included and ignored explanatory variables have positive coefficients as in our case, the result will be upward bias. Thus, we expect the omitted variable bias to cause overestimation of the contribution of the exogenous movement component if the endogenous movement component is ignored.

To test the hypothesis H1c, we quantify the linear relationship between the two movement components. Pearson's correlation coefficient of 0.8 reveals a strong and significant (p-value < 0.001) relationship between the exogenous and endogenous movement components. It means that even though they are not the same, they share a large portion of the variance. This finding alone suggests the presence of the omitted variable bias, which we further test by comparing the simultaneous model (i.e., both movement components are included) with the individual estimates (i.e., only one movement component is estimated at a time). If the coefficient estimates from the simultaneous model and the individual models significantly differ from each other, we can speak of bias.

$$simulataneous\ model\colon M_{total} = \alpha_{ex}M + \alpha_{en}M_{en} + \varepsilon_1 \quad (25)$$

$$individual\ models\colon M_{total} = \alpha_{ex}M_{ex} + \varepsilon_2\ ;\ M_{total} = \alpha_{en}M_{en} + \varepsilon_3 \quad (26)$$

The simultaneous model produces estimates for $\alpha_{ex} = 0.06$ and $\alpha_{en} = 0.35$, while the individual models arrive at $\alpha_{ex} = 0.43$ and $\alpha_{en} = 0.53$. We observe that, especially in the case of the exogenous movement, the bias results in the coefficient being 616% above the simultaneous model estimate. The bias for endogenous movement with 22% higher estimate

is less dramatic but still significant. We visualize the effect by plotting the coefficients as a regression line and comparing the slope between simultaneous and individual fit (Figure 67). Here we see that the individually fitted slope is in both cases steeper than the simultaneous.

Given these points, we confirm the hypothesis H1c assuming the presence of the omitted variable bias when coefficients are estimated individually. We found that the direct effect of urban form and allocation of activities is highly overestimated if considered in isolation.

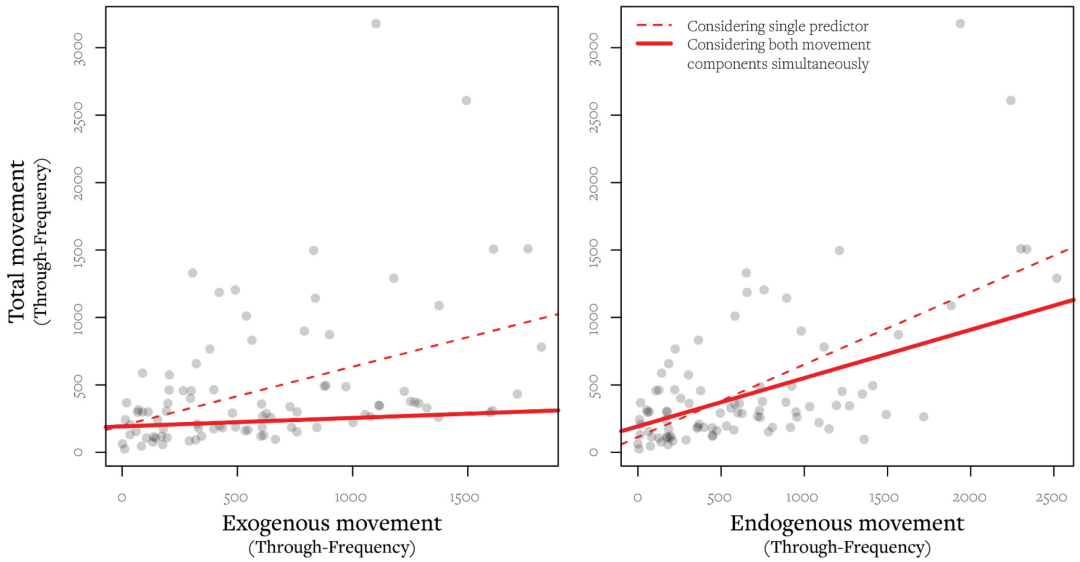


Figure 67. Regression line capturing the regression coefficient between exogenous and endogenous movement when estimated simultaneously and individually. The regression lines for each variable are constructed by holding the other variable constant at its mean. The difference in the slope represents the amount of the omitted variable bias introduced by the individual estimation.

Hypothesis H1d

Finally, we test hypothesis H1d, expecting the exogenous and endogenous movement patterns to be significantly different. Here we test for the difference in variation as well as in the amplitude.

To compare the distribution of both movement components, we run Pearson's Chi-squared test. It measures how likely it is that any observed difference between the sets arose by chance. With $\chi^2 = 29670480$ (df = 17464041) and p-value $< 2.2e^{-16}$ we refuse the null hypothesis and conclude that the two movement components are significantly different in their distribution. If we consider this finding together with the significant Pearson's correlations coefficient (r = 0.8), we can summarize that the exogenous and endogenous movement patterns are a) significantly different while at the same time b) sharing a large part of their variance.

Furthermore, we found significant differences between the amplitude of both movement components. As already discussed, the endogenous movement contributes 5.7 times more

pedestrian frequency at the 100 selected street segments than the exogenous movement. Consequently, we confirm the hypothesis H1d.

5.1.2 Movement Model Validation

After estimating the α_{ex}, α_{en} coefficients, we scale all simulated pedestrian movement components from relative to absolute pedestrian *From-Frequency, From-Volume, Through-Frequency,* and *Through-Volume* per day. By doing so, we can explore how endogenous and exogenous movement components and the individual activity types contribute to trip generation and traffic flow at any of the 7104 street segments included in the simulation.

More importantly, we are able to validate the simulated pedestrian movement by comparing it to the empirical data collected in the MiD2017. The empirical movement measured in the MiD2017 study is offering three key indices for the study area of Weimar, which can be compared to the simulated movement: *mean trip distance, mean trip volume,* and *mean trip frequency* per day. Before comparing the indices, we must normalize the simulated movement per street segment by the estimated number of inhabitants living at each street, so both the empirical and modeled movement share the same units of measurement. Since we know the total number of inhabitants for the entire study area (64855 in 2017, Statistisches Jahruch 2018) and the living space at each street segment, we calculate the estimated number of inhabitants per street segment i as follows:

$$Inhabitants_i = \frac{Living\ space_i * \sum Inhabitants}{\sum Living\ space} \tag{27}$$

Finally, we test if there is a significant difference between measured and simulated *mean trip distance, mean trip volume,* and *mean trip frequency* per day. We conclude that the two-sided unpaired t-test failed to reject the null hypothesis (see Table 2, all p values > 0.05). In other words, the simulated and empirical movement are not significantly different and confirm the validity of our model. It is important to note that a few possible factors might cause minor differences between the simulated and empirical movement. On the one hand, the simulation model captures only a fraction of endogenous movement activities (45% of all trips). On the other hand, the empirical movement *Through-Frequency* (i.e., from Weimar counting 2016, Appendix 7) used to scale the simulated movement from relative to absolute quantities also included pedestrians – mainly tourists, who are not represented in the simulation. The former is driving the estimates down, while the latter is pushing them up. In both cases, our ability to correct for the bias is limited by data availability and might be improved by future studies.

Table 2. Comparison of the summary statistics between the simulated and empirical movement.

	Simulated movement	MiD2017 Weimar	Two-sided unpaired t-test [25]
Mean Trip distance (km)	1.38	1.52	t = 1.1316, df = 2337, p-value = 0.2579
Mean Trip volume per person (km/day)	3.81	3.86	t = 0.0965, df = 2337, p-value = 0.9231
Mean Trip frequency per person (#/day)	2.74	2.53	t = 0.4092, df = 2337, p-value = 0.6824

After confirming the validity of the simulated movement components, we compare their respective patterns. Coming back to the hypothesis H1d, we visualize the variation of endogenous and exogenous movement components and explore their differences.

We do this for each movement characteristic to emphasize the different performance of each location when it comes to generating movement (i.e., *From-Movement*) as opposed to facilitating movement (i.e., *Through-Movement*). For this purpose, we present a series of figures capturing the movement distribution across all 7104 segments in the study area. All presented figures are comparable across different movement characteristics as the x-axis stays the same and depicts the individual street segments by their ID number.

[25] Only street segments with associated living space are included in the comparison.

From-Movement per Street Segment

First, we explore the movement at its origin. We found that in total, 89% of all trips (i.e., *From-Frequency*) and 83% of the travel distance (i.e., *From-Volume*) explained by the form-activity-movement interaction model can be attributed to the endogenous movement (i.e., the allocation of activities). Consequently, only 11% of all trips and 17% of travel distance are the direct product of urban form (Table 3). As seen in Figure 68 the *From-Frequency* varies strongly across the study area. The average *From-Frequency* is 25 daily trips per segment, with the maximum going up to 1309 daily trips per segment and a standard deviation of 88 daily trips. We must emphasize that a large portion of the variance can be attributed to the fact that streets with more inhabitants generate more trips than their less populated neighbors. For this reason, we also normalize the movement frequency and volume per person, as presented at the end of this section.

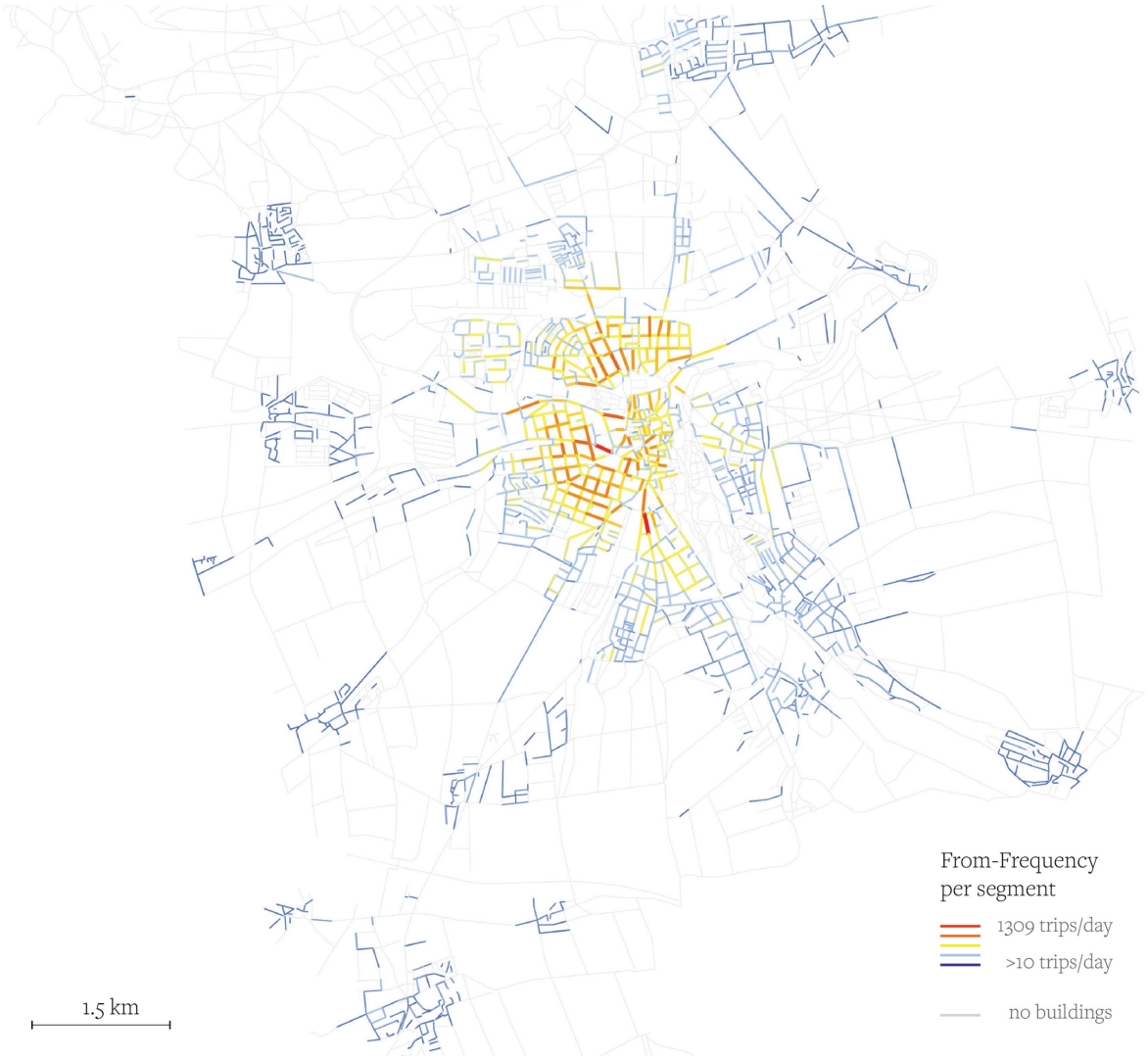


Figure 68. Spatial distribution of total pedestrian From-Frequency per day (i.e. the number of trips originating from given street segment). Red = high values, Blue = low values.

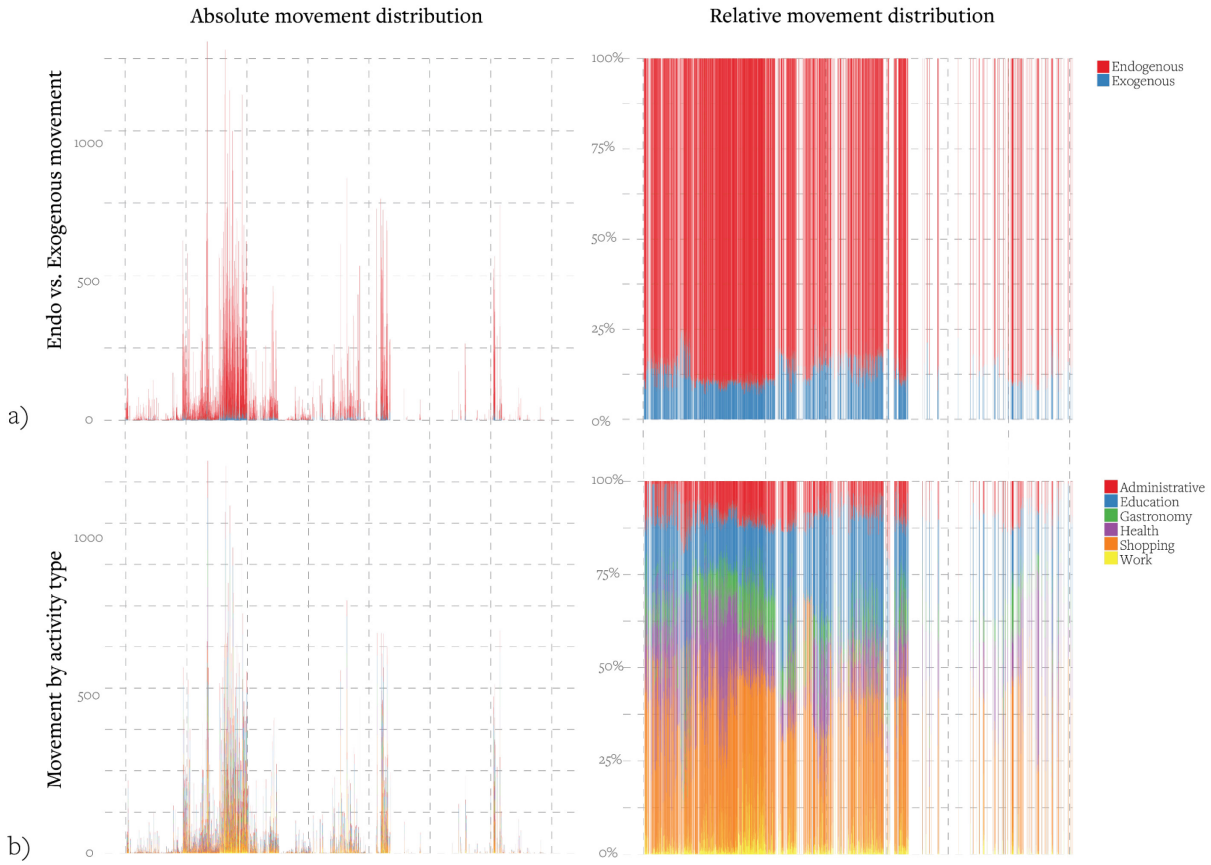


Figure 69. Distribution of total pedestrian From-Frequency per day (i.e. the number of trips originating from given street segment). a) Distribution by endogenous and endogenous movement components. b) Distribution by endogenous activity types. White spots represent segments with zero From-Frequency.

Table 3. Contribution of exogenous and endogenous movement components to the From-Movement.

		From-Frequency	From-Volume
Exogenous movement		11%	17%
Endogenous movement	Administrative	9%	11%
	Education	16%	6%
	Gastronomy	7%	8%
	Health	12%	19%
	Shopping	33%	36%
	Work	12%	3%

In the case of *From-Volume*, we observe similarly strong variation with a mean of 38 km per day[26], a maximum of 1275 km per day, and a standard deviation of 93 km per day. In contrast to the strong variation in the amplitude (Figure 70) we observe only little variation in the fraction of the total trip count attributed to each movement component (Figure 71a).

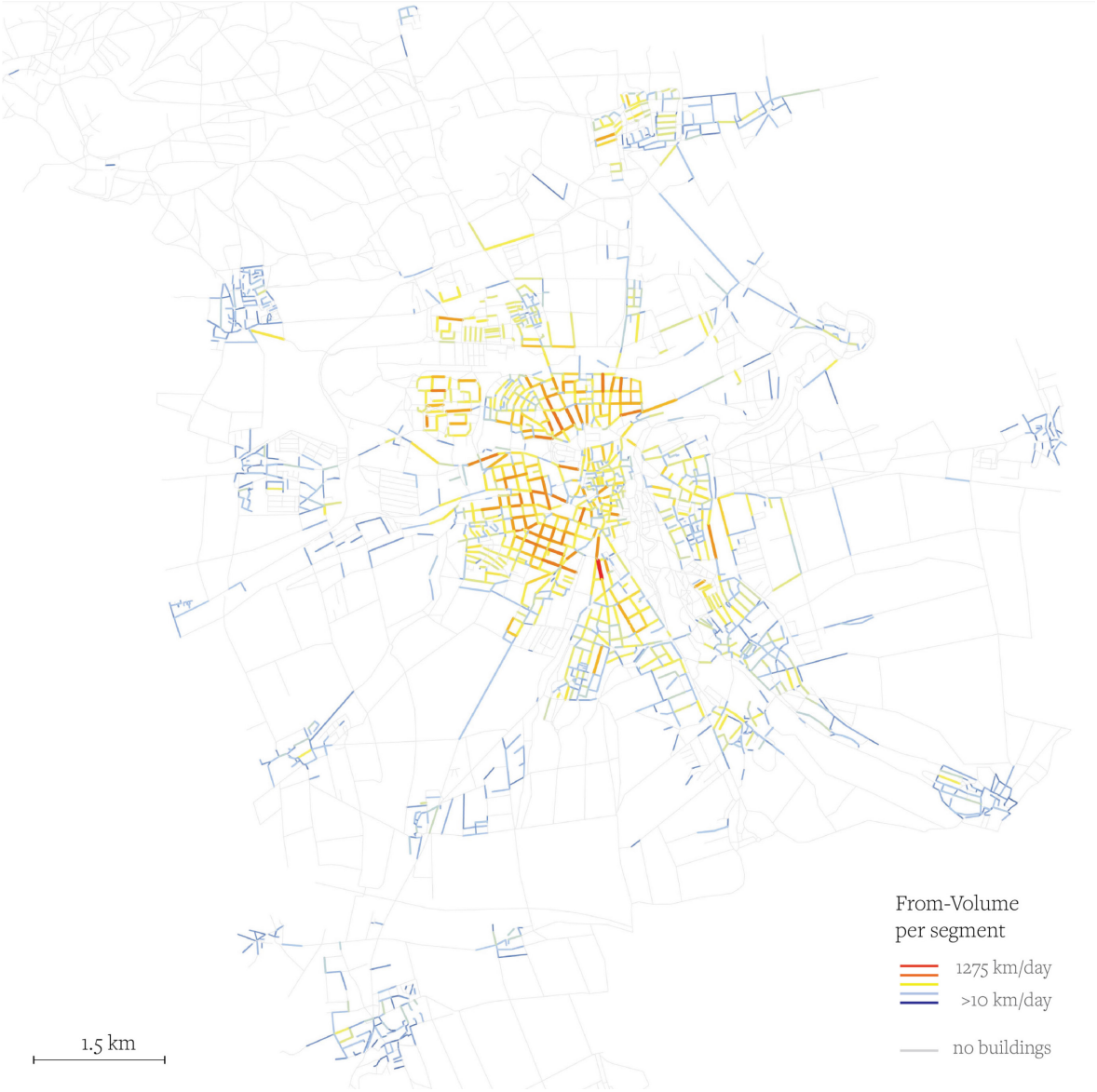


Figure 70. Spatial distribution of total pedestrian From-Volume per day (i.e., the traveled distance originating from a given street segment). Red = high values, Blue = low values

[26] The average total length of all journeys starting at any street segment. In other words, the aggregated trip length of all people living at any street.

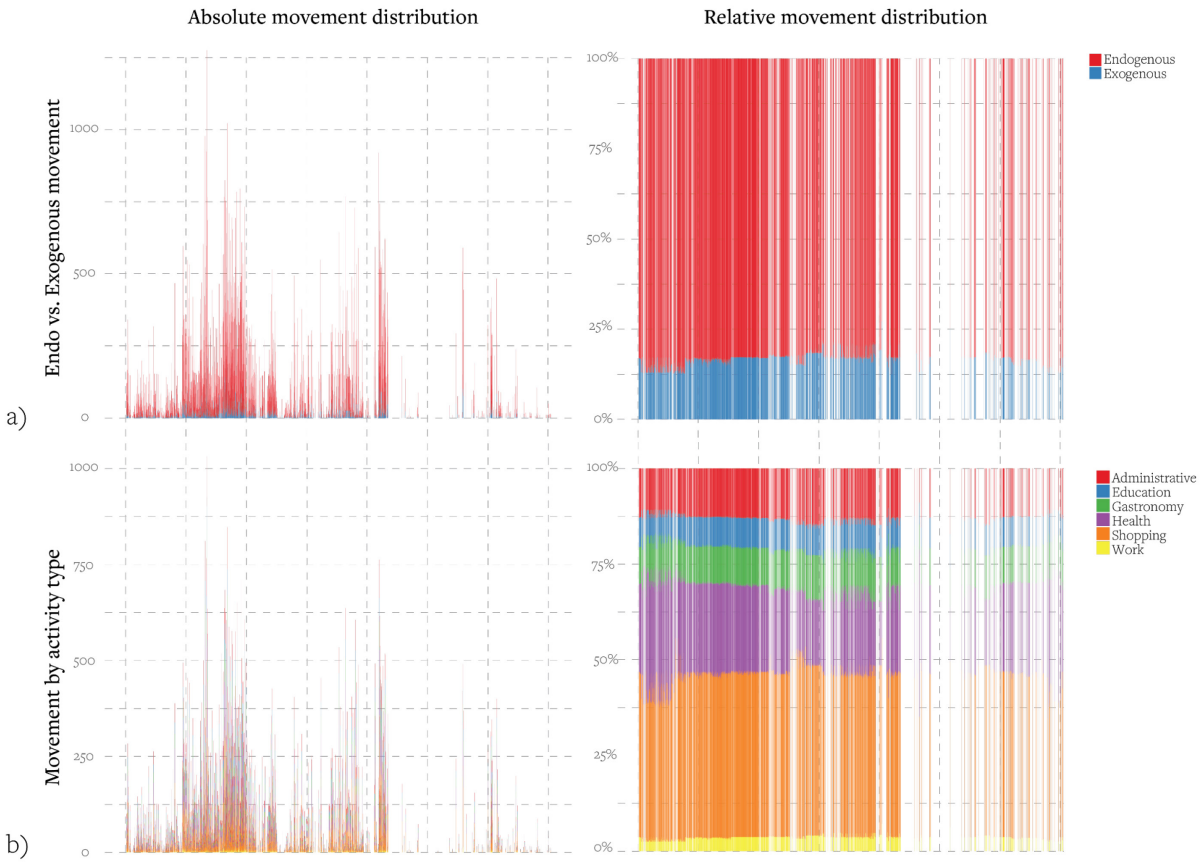


Figure 71. Distribution of total pedestrian From-Volume per day (The traveled distance originating from a given street segment). a) Distribution by endogenous and endogenous movement components. b) Distribution by endogenous activity types. White spots represent segments with zero From-Volume.

We conclude that the distribution of both, trip frequency and trip volume are closely related, as can be seen by comparing the patterns in Figure 69 and Figure 71. The strong linear association between the two patterns is underlined by the high and significant Pearson's correlation coefficient of 0.91. We attribute the strong relationship between both measures to the fact that both are highly influenced by the capacity of the street (i.e., floor area and the resulting population). Street, with more inhabitants living there, produce more trips and more traffic volume at the same time.

Through-Movement per Street Segment

As next, we explore the distribution of the passing-through pedestrian movement. We calculate the *Through-Frequency* and *Through-Volume* for each segment and observe that 83% of total Through-Volume can be attributed to the endogenous movement (Table 4). Consequently, 17% of all Through-movement is a direct product of the distribution and

configuration of urban form. Since the *Through-Volume* and the *From-Volume* equals in the sum, the overall contribution of individual movement components is also identical for both movement characteristics. Nevertheless, they differ at the level of individual street segments, as can be seen by comparing Figure 71 and Figure 75. This is confirmed by the Pearson's correlation coefficient r = 0.52 (see Table 5), suggesting that both characteristics share only 27% of their variance.

The *Through-Frequency*[27] is ranging from zero pedestrians per day up to 1935 pedestrians per day at the most frequented segment. We observe pedestrian traffic tendency to concentrate in high-frequency linear corridors often surrounded by low-frequency neighboring streets (Figure 72). Surprisingly, the most prominent pedestrian arteries circumvent the historical city center. We attribute this to the cognitive shortest path selection in the movement model resulting in the historical city center with its short, twisted street being essentially a navigational barrier.

[27] As discussed in the Methods and Data section, we do not calculate the any summary statistics (e.g. fractions, means, standard deviation) of Through-Frequency as it depends on arbitrary subdivision of the street network into segments.

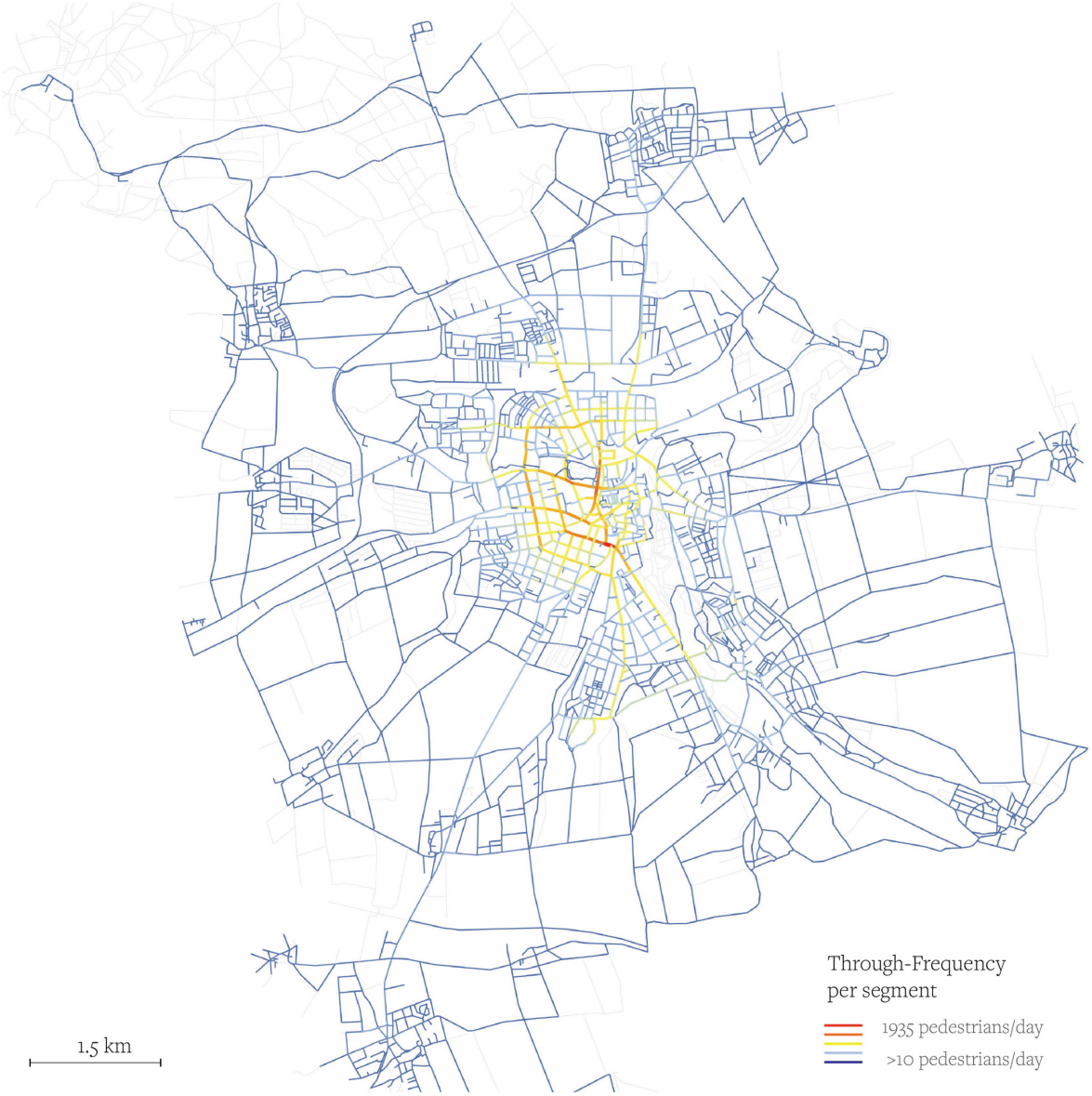


Figure 72. Spatial distribution of total pedestrian Through-Frequency per day (i.e., the number of trips passing through a given street segment). Red = high values, Blue = low values.

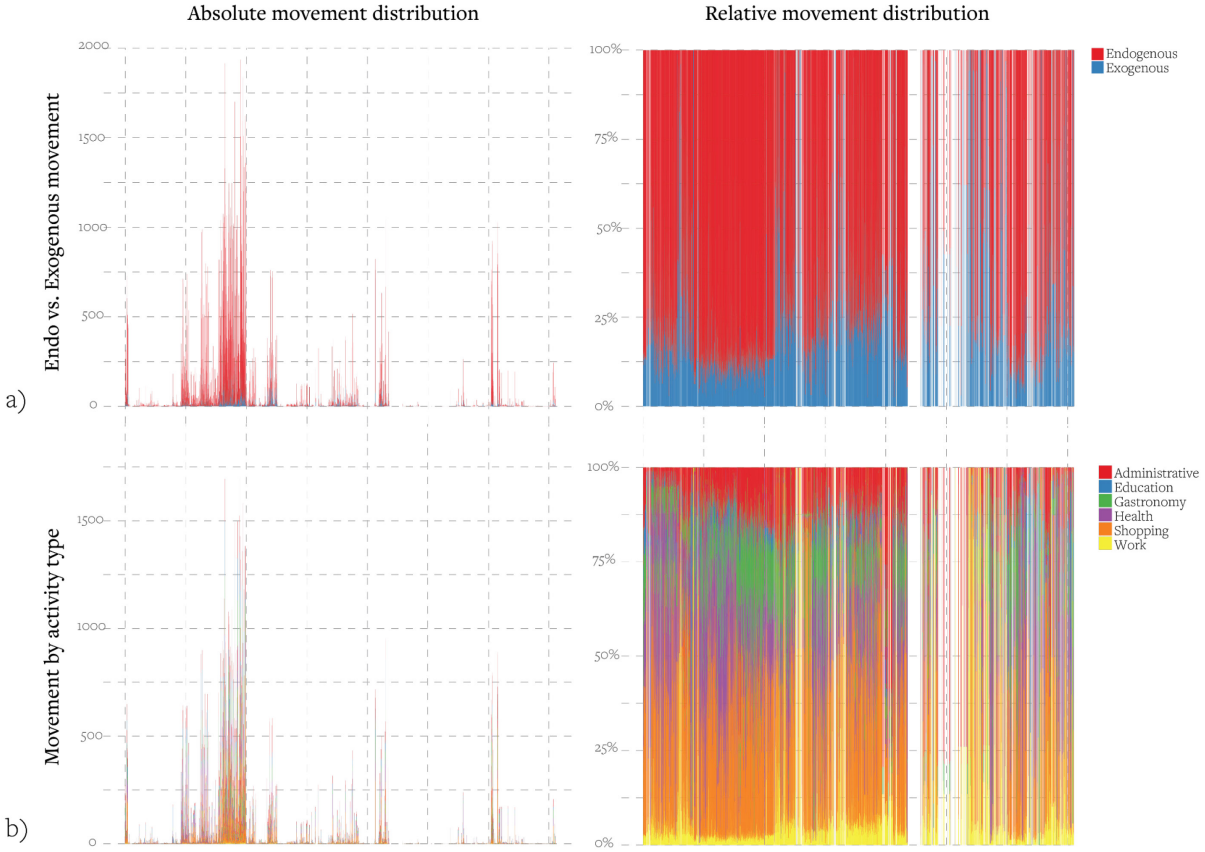


Figure 73. Distribution of total pedestrian Through- Frequency per day (i.e., the number of trips passing through a given street segment). a) Distribution by endogenous and endogenous movement components. b) Distribution by endogenous activity types. White spots represent segments with zero Through - Frequency.

In the case of *Through-Volume*, we observe the mean of 3297km of walked distance per segment, a maximum of 250980km per day, and a standard deviation of 13255 km per day. Arguably the Through-Volume is, from all four movement characteristics, the most difficult to interpret on its own. However, it is still very useful when comparing movement at a larger scale, such as neighborhoods or whole cities.

Table 4. Contribution of exogenous and endogenous movement components to the Through-movement.

		Through -Volume
Exogenous movement		17%
Endogenous movement	Administrative	11%
	Education	6%
	Gastronomy	8%
	Health	19%
	Shopping	36%
	Work	3%

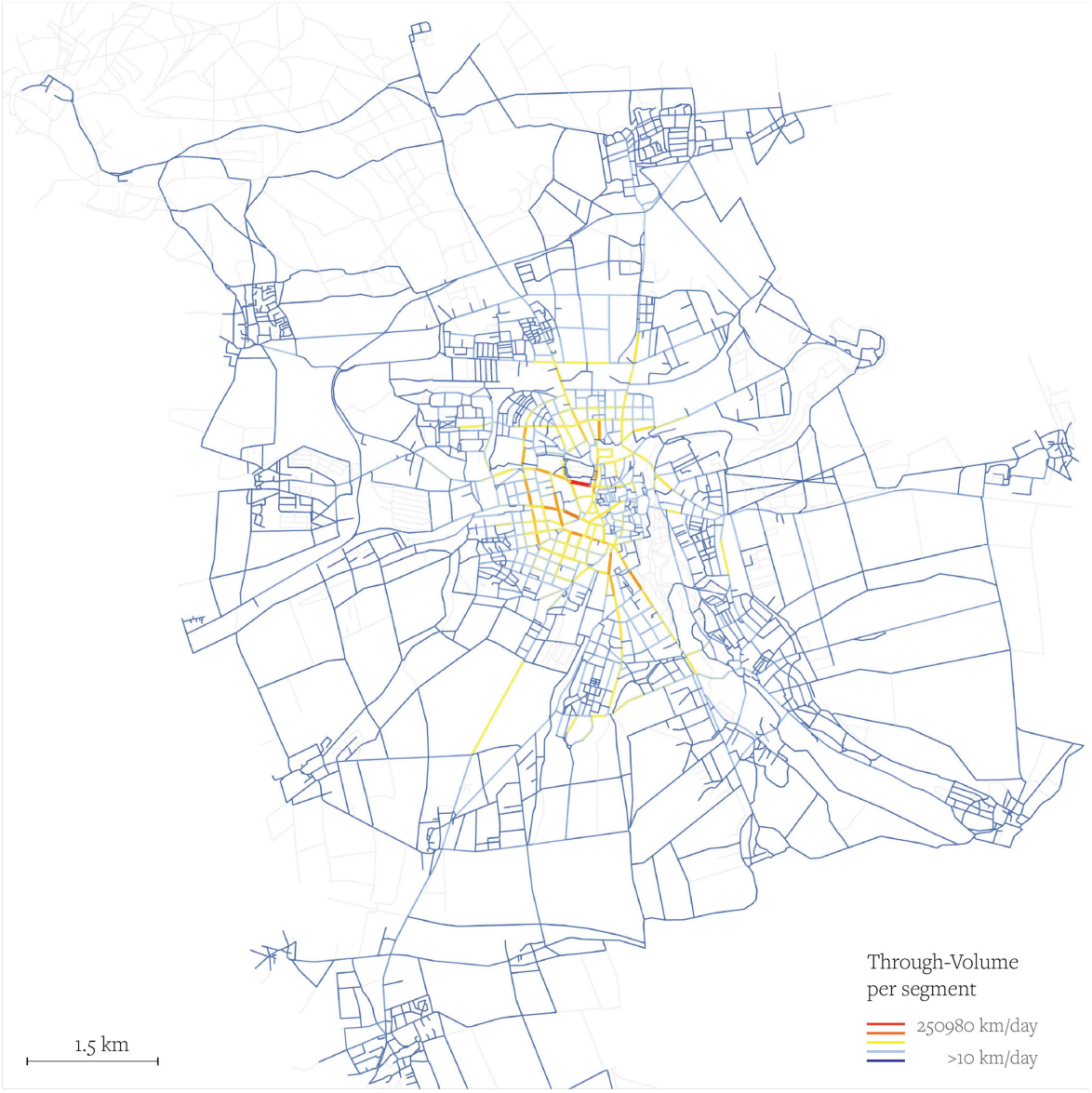


Figure 74. Spatial distribution of total pedestrian Through-Volume per day (i.e., the number of trips passing through a given street segment multiplied by its length). Red = high values, Blue = low values.

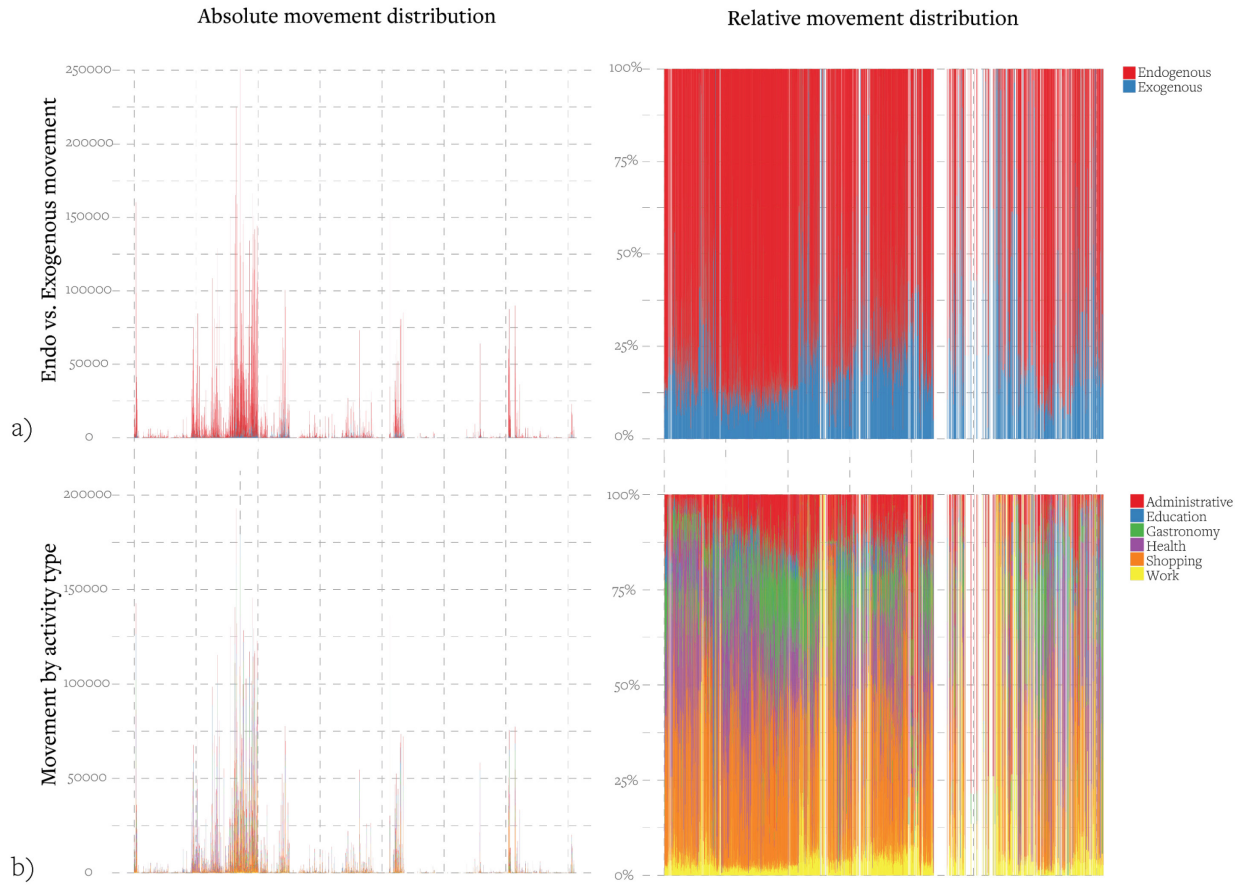


Figure 75. Distribution of total pedestrian Through- Volume per day (i.e., the number of trips passing through a given street segment multiplied by its length). a) Distribution by endogenous and endogenous movement components. b) Distribution by endogenous activity types. White spots represent segments with zero Through - Volume.

From vs. Through-Movement per Street Segment

When comparing the *From-Movement* and the *Through-Movement,* we observe the largest differences between the individual measures. As presented in Table 5, the correlation coefficient between measures coming from different categories (i.e., Through and From) ranges from 0.37 to 0.67 while the correlation between measures from within the same category goes from 0.86 to 0.91. In practical terms, it means that we often find highly frequented streets where only little movement originates and vice versa.

Table 5. Pearson's correlation matrix showing the Pearson's correlation coefficient for the four movement characteristics. Significant correlations are marked with * (p-value < 0.05).

	Through-Frequency	Through-Volume	From-Frequency
Through-Volume	0.86*		
From-Frequency	0.52*	0.67*	
From-Volume	0.37*	0.52*	0.91*

From-Movement per Person

Finally, we explore the movement distribution per person. We assess the effect a) of the urban form and b) the allocation of activities on the individual pedestrian movement. For this purpose, we normalize the movement per street segment by its estimated number of inhabitants. In the following, we discuss the *From-Movement* since it reveals how often and how much people walk based on where they live.

We found that the number of trips per person for one day is, on average, 2.74. However, based on the location, it can go up to 7.8 or be as low as 0.007[28] trips per day. The spatial distribution of trip frequency per person reveals a two-fold pattern. On the one hand, we observe gradient-like decay with the most trips being taken in the most central location, followed by a constant drop toward the edges of the study area (Figure 76). On the other hand, we observe peaks in the trip frequency at several prominent axes circumventing the city center and connecting it to the outskirts.

[28] Equals to 2.6 trips in a year.

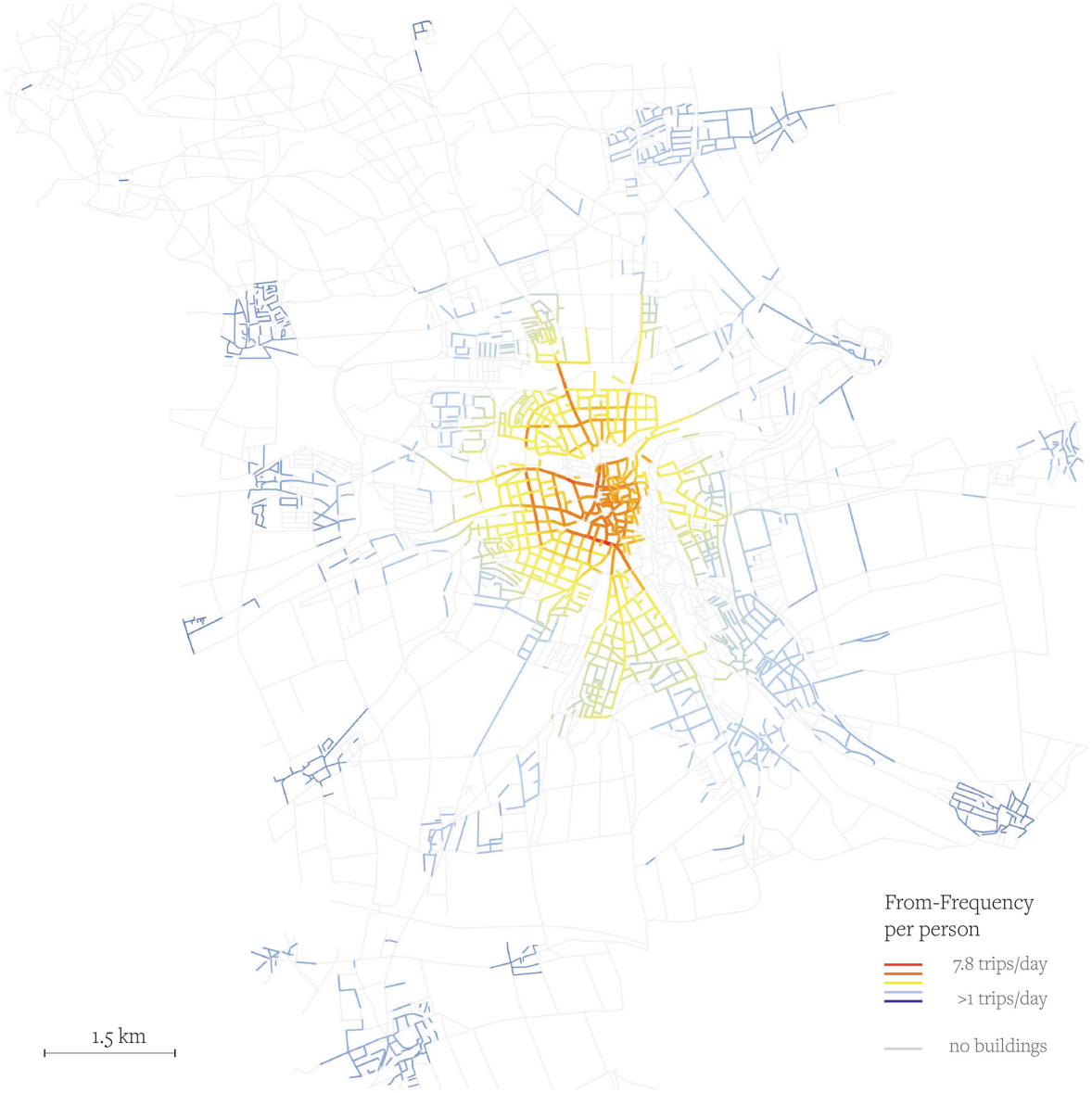


Figure 76. Spatial distribution of total pedestrian From-Frequency per person and day (i.e., the average number of trips per person living at a given street segment). Red = high values, Blue = low values.

The walking distance per person is, on average, 3.81 km per day. When compared to any other movement characteristic, the *From-Volume* per person remains relatively stable across the study area. Based on the location, it ranges from 0.121 km to 4.3km/day. Despite the discussed differences between the *From-Frequency* and *From-Volume* per person, we found a significant positive correlation between the two movement characteristics with Pearson's correlation coefficient 0.67 and p-value < 0.05. Consequently, as the number of trips per person drops from one location to another, we expect the walked distance per person to decrease. However, there are also exceptions to this general tendency – we observe locations pairs where a decrease in the trip count results in an increase in the travel distance.

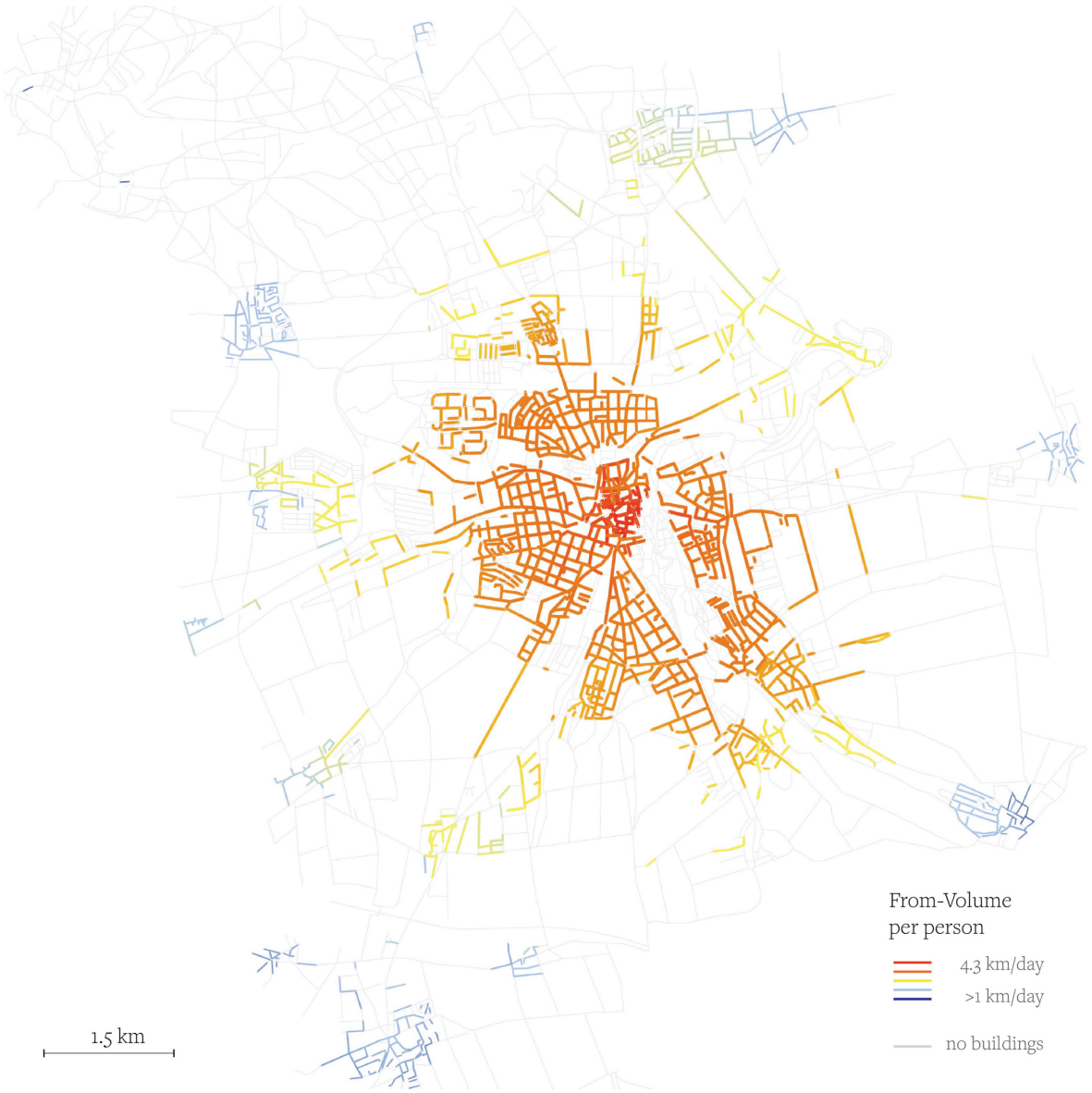


Figure 77. Spatial distribution of total pedestrian From-Volume per person and day (i.e., the average walking distance per person living at a given street segment). Red = high values, Blue = low values.

The overall spatial distribution of walking distance per person reveals a gradient-like pattern with the peak in the most central location, followed by constant decay toward the edges of the study area (Figure 77). We do not observe any local deviations from the decay pattern, and thus we conclude that the waking distance per person can be represented as a function of the distance of home from the city center. In other words, the closer a person lives to the center, the more she walks.

5.2 Activities - Testing Research Hypothesis H2

In the following, we assess the effect of endogenous and exogenous movement and autocorrelation on the allocation of activities (hypothesis H2a, H2b). Additionally, we test if the effect of both movement components must be estimated simultaneously to prevent the omitted variable bias. This is of profound practical importance since it is often the case that not both movement components are known. While the exogenous movement can be directly derived from urban form, the endogenous movement requires information on activity allocation, which might not be available when it comes to prediction of future impact and in early design stages. Under those circumstances, we can only estimate the individual effect of exogenous movement (i.e., urban form) on activity allocation. The central question is if such an estimate is not biased and, thus, if it can still be considered as useful.

The statistical model devised to estimate the effect of autocorrelation and movement on activity allocation consist of two sub-models. Each sub-model is addressing a different process driving the activity allocation. The *Filter* model identifies the critical threshold in movement required to establish an activity at a given location and to filter all observations below this threshold. The *Amplifier* takes the filtered dataset and models how variation in pedestrian movement translates into a change of activity levels. We estimate both sub-models and test the hypothesis H2 for each activity type individually.

Both statistical sub-models are accompanied by a series of tests evaluating the validity of underlying assumptions as well as graphical and written descriptions of the statistical concept. For more details, please see Appendix 17, 18, and 19.

5.2.1 Filter

The primary goal of the *Filter* is to identify and remove observations with zero-activity levels (i.e., streets without movement attractors serving only as origins of movement – only accommodation). In other words, we look for the minimum threshold in pedestrian movement required to establish different types of activities. For this purpose, we estimate a multiple binomial logistic regression model, calibrate the optimal cut-off parameter and sort out the street segments with zero-activity level.

Multiple binomial logistic regression is a statistical model belonging to the family of generalized linear models. It is used to model the probability of two discrete outcomes (e.g., “success” vs. “failure,” or in our case, street segments containing activity vs. street segments without activities). It takes multiple continuous explanatory variables (in our case, the exogenous and endogenous movement), fits sigmoid function, and returns probability for a given street to contain an activity in the range from 0 to 1 (Figure 78).

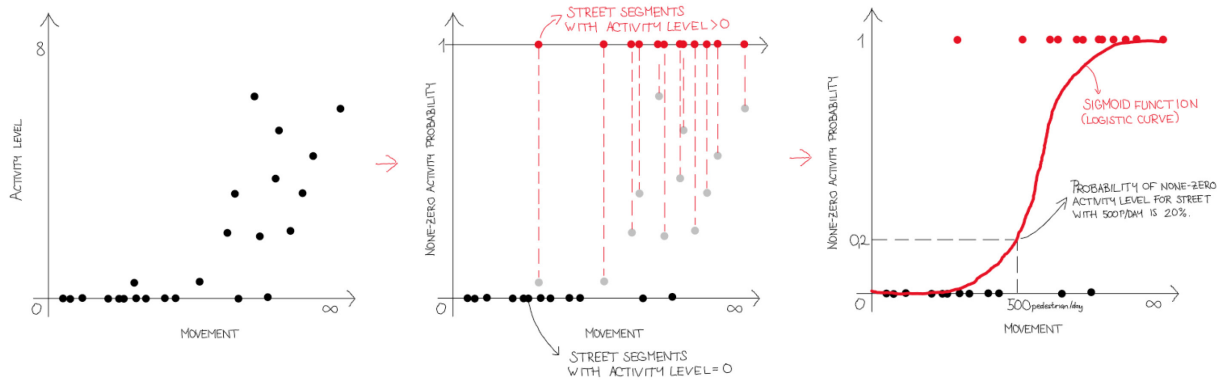


Figure 78. Simplified illustration of the binomial logistic regression with only one predictor. First, the dependent variable (i.e., activity levels) is converted from continuous to binomial scale. Next, the logistic curve is fitted through the data.

In essence, if we plug-in the value of pedestrian *Through-Frequency* at a given location, then the logistic regression model returns probabilities of finding any of the six activity types. It is important to realize that the *Filter* does not model how much, but rather if any activity can be found. We use the logistic regression model as a classifier dividing the dataset into two classes – a) zero-activity and b) non-zero activity street segments. The choice of the cut-off is essential for the ability of the classifier to recognize streets with non-zero activity levels (i.e., sensitivity) and correctly filter out streets with no activity associated with them (i.e., specificity) (for more details, see Appendix 17). It is important to realize that the cut-off level represents a trade-off between the sensitivity and specificity of the model. In our case, the ability of the regression model to filter the zero-activity observations shall not compromise its ability to recognize the streets containing activities correctly. In other words, we choose the cut-off value by prioritizing sensitivity over specificity. By doing so, we guarantee that a large portion of the none-zero activity street segments (at least 90%) remains in the dataset after the *Filter* is applied (Figure 79).

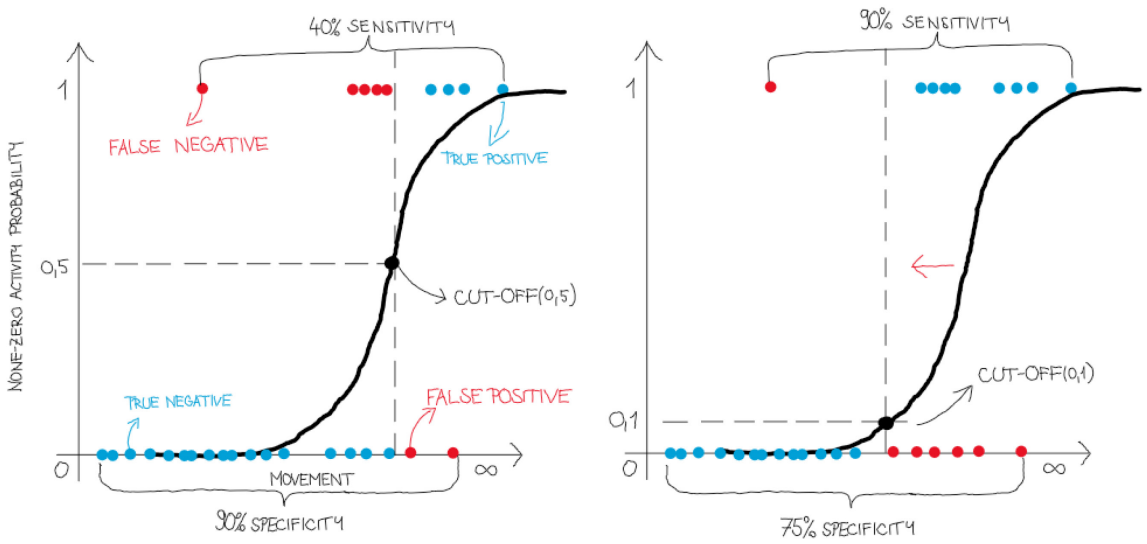


Figure 79. The impact of cut-off value on the sensitivity and specificity of logistic regression classifier (i.e., filter). We choose the optimal cut-off value to minimize misclassification error and maintain the sensitivity above the 90% threshold.

In the following, we present the summary of six logistic regression models (i.e., one for each activity type) and discuss the contribution of the individual movement components and the overall ability of each model to filter out street segments with zero-activity level. For a detailed report, including the estimated model coefficients and performance metrics, please see Appendix 17.

We found the pedestrian movement to be a significant predictor of zero-activity street segments. The *Filter* was particularly efficient in the case of the shopping, gastronomy, and health activity as it was able to remove 83%, 64%, and 55% of respective zero-activity street segments (Table 6). For all three activity types, both movement components were significant predictors of zero-activity observations. For the work, education, and administrative activities, the *Filter* performance drops sharply to 34%, 30%, and 27% specificity level. For these three activity types, a large portion of zero-activity street segments remains in the dataset, which presents a difficulty for the linear regression model employed by the *Amplifier* model.

Table 6. Performance of the Filer model showing specificity (i.e., percentage of correctly filtered zero-activity streets) and significance of endogenous and exogenous movement components (* means significant with p-value < 0.05, / means not significant with p-value > 0.05).

	Administrative	Education	Gastronomy	Health	Shopping	Work
Sensitivity	94%	90%	90%	92%	90%	90%
Specificity	27%	30%	64%	55%	83%	34%
Exogenous movement	/	*	*	*	*	*
Endogenous movement	*	/	*	*	*	*

To illustrate the different performances of the *Filter* based on the activity type, we present a graphical comparison of the best and worst-performing model (all models can be found in Appendix 17). We visualize the logistic curve for both predictors (i.e., endogenous and exogenous movement) and highlight true and false model classifications (Figure 80). In the case of shopping activity, we observe a sharp rise in the probability of finding activities (i.e., the inclination of the logistic curve) as the movement level increase. As a result, we were able to effectively filter out 83% of the zero-activity street segments while keeping 90% of non-zero activity segments. However, when it comes to administrative activities, we observe a much shallower logistic curve resulting in the overall classification accuracy of only 29%. Moreover, we found that the exogenous movement, which can be directly derived from the urban form, was not a significant predictor of zero-level administrative activities.

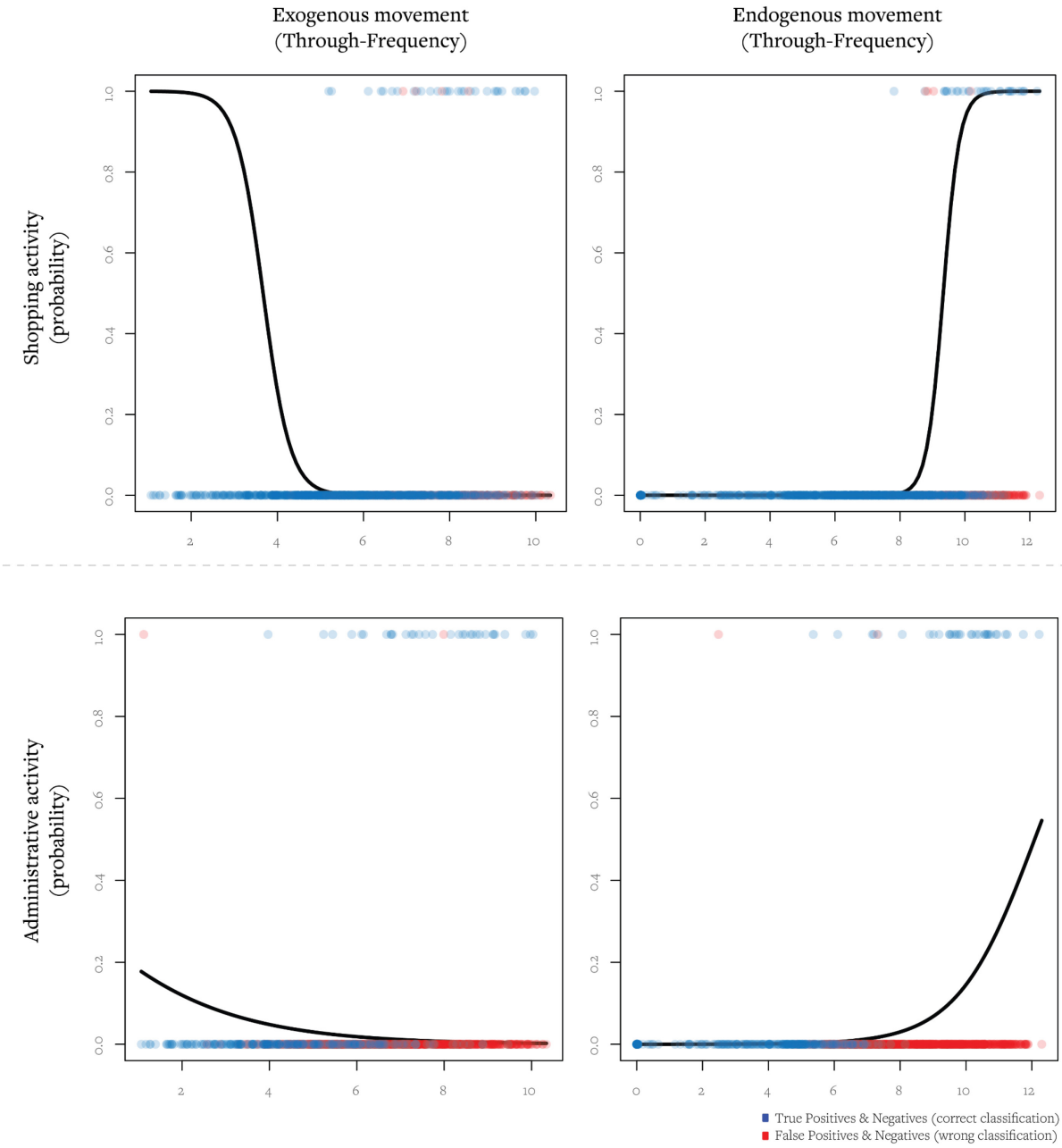


Figure 80. Comparing the best and worst performing logistic regression model. We visualize the logistic curve for both explanatory variables. The displayed curves for each variable are constructed by holding the other variable constant at its mean value. Correctly classified observations are marked blue, misclassified observations are marked red.

5.2.2 Amplifier

After using the *Filter* to select the street segments with activities associated with them, we model how the change in pedestrian movement affects the intensity of these activities. We expect the activities to be either attracted or repelled by movement and formally represent this relationship through the linear regression (LM) model. Moreover, we expect the activity intensity at any location to be influenced by the activity intensity at the neighboring locations. For this reason, we extend the LM model by spatial autoregressive term accounting for spatial dependence between activities.

As discussed in Chapter 4.3.2 and Appendix 18, the resulting spatial autoregressive model (SARLM) is formally defined as:

$$LM: y = \alpha X + \mu \tag{28}$$

$$SARLM:\ y = \rho W y + \alpha X + \mu \tag{29}$$

or in our case,

$$activity_T = \rho W activity_T + \alpha_{ex} m_{ex} + \alpha_{en} m_{en} + \mu \tag{30}$$

The $activity_T$ is a vector that is capturing the intensity of activity type T, W stands for spatial weights matrix, ρ is a spatial auto-regressive coefficient and m_{en}, m_{ex} are explanatory variables capturing the exogenous and endogenous movement volume at a given location [29].

As it is the case in any type of statistical modeling, we simplify complex reality into models that aim to be useful tools for answering a particular question. In general, the best model is the simplest one which does the job. Every extension of the model makes the model more difficult to interpret, which should be evaluated against the possible gains in precision. As a result, the process of building the statistical model follows *Occam's razor* principle, which suggests starting with the simplest model possible and extending it only if necessary. Accordingly, we begin with simple linear regression, perform a series of spatial and non-spatial specification tests, and based on the results, we either stop or introduce the spatial autoregressive term $\rho W activity_T$.

We estimate the resulting regression model to test the research hypothesis H2 about the effect of spatial autocorrelation, exogenous movement, and endogenous movement on the allocation of activities. Like in the case of testing the hypothesis H1, we test if the effect of both movement components must be estimated simultaneously in order to prevent the omitted variable bias.

[29] Throughout the chapter 5.2, we adopt a matrix notation. The vector variables are marked with lowercase letters and matrix variables wit upper case letters.

In the following, we present a brief summary of the *Amplifier* model and results of the hypothesis test H2a, H2b, H2c, and H2d for all six activity types. For detailed documentation and specification of the model building process for all activity types, please refer to Appendix 18.

Hypothesis H2a and H2b

We found the pedestrian movement to be a significant predictor of activity intensity in the case of three out of six activity types. In the case of gastronomy, shopping, and work activities, the linear regression model was able to explain 66%, 58%, and 41% of the respective variance in activity intensity (Table 7). When looking at the individual movement components, the endogenous movement coming from the activities themselves was a significant predictor for all three activity types. The exogenous movement directly derived from the urban form was a significant predictor only in the case of the work activity.

Overall, we found movement being a significant predictor of the presence of activities (i.e., *Filter* model); however, it has only limited ability to explain the variation in activity intensity (i.e., *Amplifier* model). In the case of exogenous movement being directly derived from the urban form, we found that five out of six activities were significant at least in one of the two sub-models (i.e., *Filter* and *Amplifier).* As a result, we reject the hypothesis H2a in case of administrative activity and confirm it for the rest. The endogenous movement was also a significant predictor of five activity types except for education. Consequently, we confirm the hypothesis H2b in five cases and reject it in one case.

Table 7. Summary table of the linear regression model with exogenous and endogenous movement as explanatory and activity intensity as dependent variables. Significant explanatory variables are marked with ***.

	Administrative	Education	Gastronomy	Health	Shopping	Work
R^2	/	/	0.66	/	0.58	0.41
Exogenous movement	/	/	/	/	/	***
Endogenous movement	/	/	***	/	***	***

Hypothesis H2d

We tested all three significant models identified in the previous step (i.e., gastronomy, shopping, and work) for the omitted variable bias. The analysis of variance (ANOVA) Chi-squared test comparing the simultaneous and individual models proves a significant difference ($p < 0.05$) between the simultaneous and individual models (Appendix 18). This means that in all three cases, we confirm the hypothesis H2d.

In Figure 81, we show the impact of the omitted variable bias on the strength and direction of the relationship between movement components and activity intensity. We observe the two-fold effect of the omitted variable bias. When considering the exogenous movement (i.e., effect of the urban form) and endogenous movement (i.e., coming from activities themselves) in isolation, the estimated effect a) points in the wrong direction[30] or b) non-significant explanatory variable is falsely identified as significant[31].

[30] Exogenous movement in case of shopping and endogenous movement in case of work activities

[31] Exogenous movement in case of gastronomy.

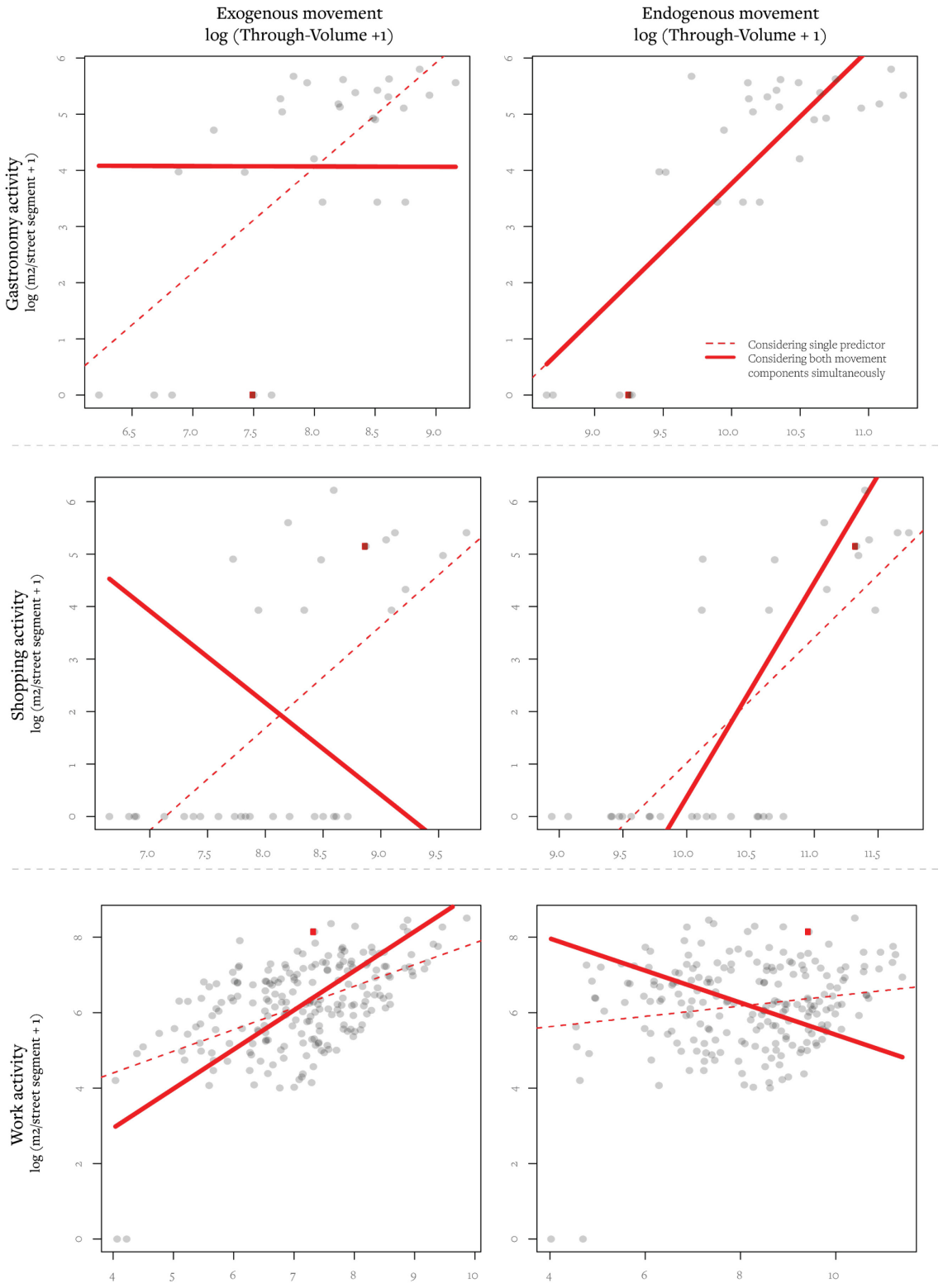


Figure 81. Regression line capturing the relationship between exogenous and endogenous movement and activity intensity when estimated simultaneously and individually. The regression lines for each variable are constructed by holding the other variable constant at its mean. The difference in the slope represent the amount of the omitted variable bias introduced by the individual estimation.

Hypothesis H2d

To test for the spatial dependence between activities, we run a series of spatial diagnostic tests (Morans's I, Lagrange multiplier). Without going into details (see Appendix 18), these tests are designed to reject the null hypothesis of no spatial autocorrelation (i.e., diffuse test) or to identify the type of spatial dependence (i.e., focused test).

In the case of gastronomy and work activities, the test results suggest the presence of spatial autocorrelation ($p < 0.05$). Hence, we run a spatial autoregressive model for both activity types. On the one hand, we found a negative spatial autoregressive coefficient for gastronomy, while, on the other hand, the coefficient was positive for the work activities. In essence, this means that we observe clustering of the work and dispersion of the gastronomy activities throughout the study area. It is important to realize that the choice of the spatial measurement unit, in our case, the street segment, plays a crucial role in the strength and direction of the spatial dependence. It is possible that gastronomy is clustering on a finer spatial resolution (e.g., building), but when it comes to streets, we observe that a high concentration of gastronomy on a given street negatively influences its intensity in the neighboring streets.

We quantify the indirect effect of spatial autocorrelation and conclude that the effect of pedestrian movement on gastronomy gets reduced by 49% due to the negative spillovers. On the contrary, the positive spillovers amplify by 7% the effect of pedestrian movement on work activities. In the case of shopping, no spatial spillovers could be detected (Table 8).

To summarize, we confirm the hypothesis H2d in case of gastronomy and work and refuse it in case of shopping activities. For the remaining three activity types – administrative, education, health, we were not able to test the hypothesis H2d.

Table 8. Indirect effect of spatial dependence.

	Administrative	Education	Gastronomy	Health	Shopping	Work
Spatial dependence (indirect effect)	/	/	-49%	/	NO	7.19%

5.2.3 Activity Allocation Model Summary

After estimating the *Filter* and *Amplifier* sub-models, we explain the allocation of activities as a function of exogenous and endogenous pedestrian movement and spatial autocorrelation. We run the prediction model for all three activity types significantly related to pedestrian movement and calculate the residual mean and standard error of the prediction (for details, see Appendix 19). In simple terms, we measure the difference between the predicted and actual activity intensity at each street segment (i.e., residual) and calculate the average relative deviation of the predicted activity intensity (in m^2 of floor area). We

found the predicted shopping and gastronomy to be on average no more than 0.25% and 1.18% different from the actual value, while the prediction of work activities is with the mean error of 14% percent less accurate (Table 9). The high prediction accuracy can be attributed to the fact that most streets have a zero-activity intensity, and the ability of the *Filter* to correctly identify them dramatically increases the performance of the combined activity prediction model.

Table 9. Accuracy of the activity prediction model (Filter + Amplifier) based on the exogenous and endogenous movement.

	Administrative	Education	Gastronomy	Health	Shopping	Work
Relative mean error (%)	/	/	1.18%	/	0.25%	14.20%

When looking at the spatial distribution of the estimated activities, we specifically focus on the model residuals (e.g., the prediction errors). In essence, these are showing the difference between the actual activity intensity and the predicted activity potential given by pedestrian movement (see Appendix 19 for more details). Positive residuals represent locations with actual activity intensity being higher than the predicted, and the negative residuals show the opposite. We find both, the positive and negative residuals, to be spread across the whole study area in case of work activities (Figure 84) and concentrated around the city center in case of shopping (Figure 83) and gastronomy (Figure 82). For all three activities, we found the negative residuals to be concentrated in the residential belt around the historical center. On the contrary, the positive residuals are dispersed all over the study area building small clusters of neighboring streets.

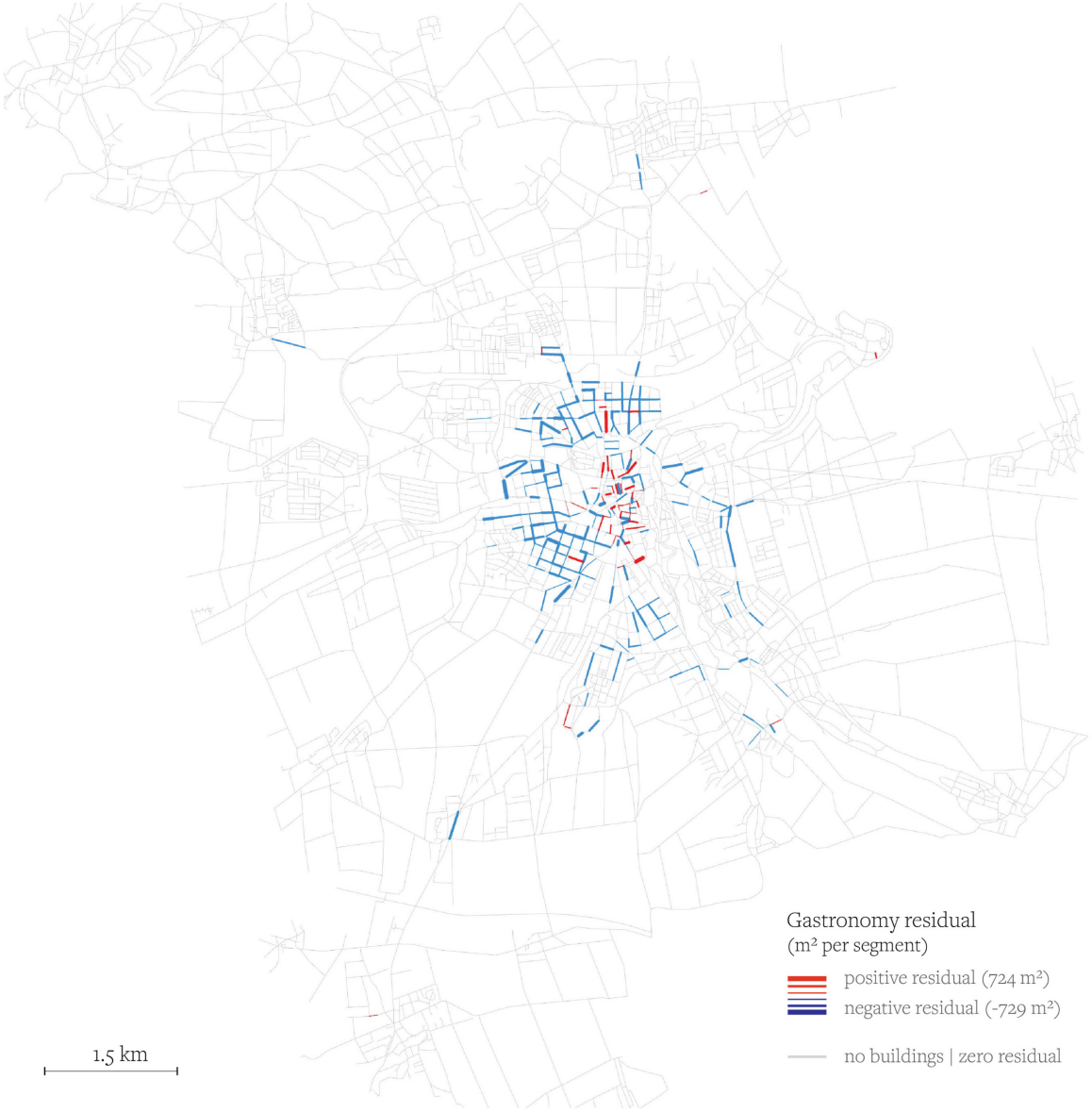


Figure 82. Residual map for predicted intensity of gastronomy activity. (Blue = negative residuals, Red = positive residuals, Black = prediction on target. The line thickness expresses the amplitude of the residual).

Figure 83. Residual map for predicted intensity of shopping activity. (Blue = negative residuals, Red = positive residuals, Black = prediction on target. The line thickness expresses the amplitude of the residual)

Figure 84. Residual map for the predicted intensity of work activity. (Blue = negative residuals, Red = positive residuals, Black = prediction on target. The line thickness expresses the amplitude of the residual).

We found a significant effect of endogenous and exogenous activity allocation on pedestrian movement and showed that both movement components differ in their contribution to total movement as well as in their spatial pattern. Moreover, we demonstrated that estimating the effect of exogenous activities on movement in isolation (i.e., without the effect of endogenous activities) leads to bias.

5.3 Hypotheses Test Overview

To summarize, we estimated the form-activity-movement interaction model to test the bias of the CUM model explaining the pure effect of urban form on pedestrian movement and allocation of activities. For this purpose, we tested research hypotheses about:

a) the effect of movement and spatial autocorrelation on activities,
b) the effect of activities and urban form on movement, and
c) about omitted variable bias caused when the above-mentioned effects are estimated in isolation.

Table 10. Testing hypothesis 1.

Q1	What is the effect of urban form and allocation of activities on the pedestrian movement?	
	Hypothesis	Result
H1a	Pedestrian movement is directly affected by urban form.	Accepted
H1b	Pedestrian movement is affected by the allocation of activities	Accepted
H1c	The direct effect of urban form and the effect of activities on movement must be estimated simultaneously. Individual estimation is a source of bias.	Accepted
H1d	The pedestrian movement pattern directly generated by urban form and the pedestrian movement pattern directly generated by the allocation of activities are significantly different.	Accepted

Table 11. Testing hypothesis 2.

Q2	How do the urban form, pedestrian movement, and spatial autocorrelation affect the allocation of activities?						
	Hypothesis	Result by activity type					
		Administrative	Education	Gastronomy	Health	Shopping	Work
H2a	Allocation of Activities is affected by exogenous pedestrian movement generated by the configuration of urban form.	Rejected	Accepted	Accepted	Accepted	Accepted	Accepted
H2b	Allocation of Activities is affected by the endogenous pedestrian movement generated by themselves.	Accepted	Rejected	Accepted	Accepted	Accepted	Accepted
H2c	Activities are spatially autocorrelated. This means that activities at a given location are affected by activities at the neighboring locations.	NA	NA	Accepted	NA	Rejected	Accepted
H2d	Ignoring any of the effects described in H1a, H1b, and H1c cause omitted variable bias.	NA	NA	Accepted	NA	Accepted	Accepted

6 Discussion and Conclusions

When Socrates famously said that "by far the greatest and most admirable form of wisdom is that needed to plan and beautify cities and human communities" he merely suggested that planning and building cities is not an easy task. On the contrary, the sheer complexity of the task makes it notoriously difficult and requires the coordinated effort of many professionals coming from different disciplines. Based on their background, they provide different perspectives on how urban form affects those living in cities.

In this study, we confronted the approach of configurational urban morphologists (CUM), transportation planners (TP), and urban economist (UE) when it comes to modeling the interaction between urban form, human activities, and pedestrian movement. After reviewing the literature on each discipline, we found three distinct puzzle pieces and proposed a unifying model bringing them all together. In short, TP scholars focus predominantly on movement and consider it as a product of activity allocation. The UE focuses primarily on activities and explains their distribution through their interactions. And finally, the CUM scholars aim their attention on the urban form and interpret the movement and activity allocation as an effect of its configurational properties.

We argue that the approaches followed by each discipline are complementary, and thus, instead of comparing them against each other, we devised a joined interaction model with the feedback-loop between a) activities and movement and b) activities themselves, both interacting through and with the urban form. To unify the different concepts about where movement comes from, we split the variation in activities and movement into the exogenous and endogenous components. By doing so, we were able to accommodate the idea that movement can be directly derived from urban form together with the approach arguing that movement is driven by the allocation of activities. In essence, both propositions are not contradictory since each of them is explaining a different component of movement and activity allocation pattern.

By bringing the individual interactions identified during the literature review into the joined form-activity-movement interaction (FAMI) model, we revealed a system of simultaneous interactions and its possible consequence for the individual approaches of CUM, TP, and UE. Since they estimate the effect of individual interactions between form-activity-movement in isolation ignoring their simultaneity, we raised the question of their validity. On the one hand, we acknowledge the benefits of simple, focused models, on the other hand, we argue that we must be aware of the trade-off between their simplicity and accuracy. Moreover, by acknowledging that "all models are wrong, but some are useful"(Box, 1979, p202), the question is not which of the three models is the most precise one, but rather if their usefulness has not been compromitted for the sake of simplicity.

Since estimating the effect of some interactions while ignoring others is notoriously known as the source of the omitted variable bias, we argue that the theoretical arguments behind the individual models must undergo the scrutiny of the empirical test. As Richard P. Feynman puts it in his 1964 lecture, “It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong”.

In our view, it is the empirical experiment and hypothesis testing, which presents the most significant challenge and contribution of this study. The most prominent source of difficulties is the plain fact that despite the theoretical plausibility of splitting the joined FAMI model variables into the exogenous and endogenous components, there is no established way how to measure it empirically. Since there is no such thing as exogenous pedestrians or endogenous shop owners, a large part of the model variables could not be directly measured and had to be estimated, simulated, and derived from the joined distribution of activity allocation and pedestrian movement. For this purpose, we collected and acquired empirical data capturing the pedestrian movement flows, route choice behavior, travel diaries, and distribution of travel activities in the mid-size German city of Weimar.

This data was used to test the research hypothesis and to calibrate the pedestrian movement simulation model. Since the distribution of the exogenous and endogenous movement could not be directly measured, we estimated its properties through simulation. Prior to running the movement simulation, we conducted a series of studies to calibrate the pedestrian movement model parameters. We selected six travel activity types and tested three alternative approaches for aggregating them from buildings to streets. We compared two established route-choice models and found the cognitive model based on angular shortest paths as a more accurate representation of Weimar's pedestrian movement. Finally, we used the travel diaries to calibrate the travel impedance function representing the willingness to walk in relationship to the distance and type of the destination. Another key point to mention is our emphasis on the interpretability of the simulation results. We investigated the variation of movement throughout the daytime and weekdays to identify the smallest representative temporal unit of measurement.

Additionally, we introduced the concept of movement as a multi-dimensional phenomena described four distinct movement characteristics: From-Frequency, From-Volume, Through-Frequency, and Through-Volume. We discussed their theoretical properties and simulated their distribution patterns across the study area. The simulated movement characteristics reveal how often and how far people walk based on where they live and how the movement flows are distributed across space.

After simulating the exogenous and endogenous movement components, we test the hypothesis H1 about the effect of a) urban form and b) allocation of activities on pedestrian

movement. We estimate the effect of both explanatory variables simultaneously and in isolation, which allows us to test for the omitted variable bias. As next, we test hypothesis H2 about the impact of a) pedestrian movement, b) urban form, and c) spatial autocorrelation on activities. Similar to the hypothesis H1, we compare the outcome of simultaneous estimation with the individual approach and test the for the possible source of bias.

For the purpose of the hypothesis testing, we employed a set of regression models specifically selected to address the unique properties of pedestrian movement and activity allocation. When it comes to movement, we employed a penalized regression model to guarantee that only non-negative movement estimates are allowed. This was necessary to ensure the theoretical validity of the statistical model as a negative movement is conceptually not feasible. In the case of the activities, we devised a novel two-stage regression model addressing their spatial autocorrelation and non-normal right-skewed bi-modal distribution.

6.1 Discussion of Results

To start with, we must note that all findings presented in this study are constrained to the city of Weimar, Germany, and provide no causal or definite evidence. The observational character of the study design does not allow for generalizations as "no number of sightings of white swans can prove the theory that all swans are white." Popper (1959).

With this in mind, we summarize the results of the hypothesis tests about the interaction between activities, movement, and urban form. We start by discussing the effect of urban form and activity allocation on pedestrian movement. Afterward, we examine the same relationship in the opposite direction.

6.1.1 Effect of Urban Form and Activities on Movement

We found both, the activity allocation and urban form to be significant predictors of pedestrian movement, explaining 58% of its variance. Regarding their contribution to the overall movement, we found that the activities account for 89% of the explained trips and 83% of the travel distance. With the direct effect of urban form accounting for the remaining 11% of trips and 17% of travel distance, we found the allocation of activities to be a major determinant of pedestrian movement.

The unexplained 42% of the variance in pedestrian movement shall be attributed to a) activity types ignored by our model and b) pedestrian behavior not covered by our model (e.g., tourists, chain-trips). On the one hand, the fact that we ignored a large portion of travel activities might cause that the true portion of movement explained as the direct effect of urban form is most likely below our current estimate.

These findings confirm the fundamental claims made by transportation planners and configuration urban morphologist as both, the activities and the direct effect of urban form were significant sources of pedestrian movement. What might be surprising and contradictory to the previous findings is their relative contribution. Repeated claims Space Syntax scholars suggesting that "over 60% of human movement can be predicted or explained purely from a topological point of view" (Lerman et al., 2014, p395) stand in sharp contrast to our findings. The configurational urban morphologists routinely estimate the "pure" effect of urban configuration on movement and often come to results suggesting that the majority of movement can be directly attributed to urban form only (Hillier et al. 1993; Penn et al. 1998; Hillier and Iida 2005; Jiang 2009a, Read 1999). The results of our study suggest that the explanation of this apparent contradiction might be the bias of the "pure" model considering only the effect of urban form. We found that if the effect of activities or urban form is estimated in isolation, the model is upward biased. In our case study, the direct effect of urban form was overestimated by 616% when we omitted the impact of activities. In essence, a portion of the ignored effect of activities gets wrongly attributed to the urban form rendering the whole model as flawed.

The key conclusion here is that even though the direct effect[32] of urban form alone explains a significant portion of pedestrian movement, we must also consider the allocation of activities to estimate its contribution correctly. Otherwise, we might be wrongly attributing too much weight to the direct effect of urban form at the expense of activities. We conclude that such a model might be misleading when used to inform planners and policymakers. We are aware of the fact that in cases such as the early design stages when the information on activity allocation is simply not available, estimating the "pure" effect of urban form in isolation might be the only option. However, we argue that the analyst must be cautious when interpreting the results as these are most likely overestimating the direct effect of urban form.

6.1.2 Effect of Movement, Urban Form, and Spatial Autocorrelation on Activities

When it comes to explaining the allocation of activities, we found that pedestrian movement is a significant predictor of all activity types. In the case of gastronomy, shopping, health, and work activities both, the exogenous movement generated by urban form and the endogenous movement generated by activities themselves were significant predictors of their

[32] We must note that we specifically estimate the direct effect of urban form, since there is also an indirect effect of urban form propagating through the feedback loop between movement and activities. In the current study we did not measure this indirect effect as it is irrelevant in context of testing the validity of the existing approaches. Since the CUM approach focuses only on the direct impact of urban form and the TP and UE approaches are not concerned with the effect of urban form at all we limited our methodology to the estimation of direct effects of urban form on movement.

allocation. Moreover, we found the exogenous movement generated by the urban form being unrelated to administrative activities and the only significant predictor of educational activities.

We must note that for three out of the six activity types – administrative, education, and health activities, the pedestrian movement could only predict if they are present, but not at which intensity. In the case of the remaining gastronomy, shopping, and work activities, the pedestrian movement could explain 41- 66% of the variance in their intensity.

It is important to say that estimation of the effect of pedestrian movement on activity allocation requires both, the endogenous and exogenous movement components in order to prevent the omitted variable bias. We found that if the effect of exogenous movement is estimated in isolation, the estimates are not only less accurate but also pointing in the wrong direction. Such biased forecasts might have a detrimental impact on urban development as the planning decisions and policies might lead to opposite effects as intended.

However, when it comes to predicting the future allocation of activities, we only have information on the exogenous movement generated by the urban form only. The endogenous movement component is in such cases unknown as it is driven by the allocation of activities which is being predicted. As a result, these predictions are prone to the omitted variable bias and thus do not qualify as reliable.

Despite of this limitation, we argue that the form-activity-movement interaction model can still be useful for the analysis of existing urban environments as it reveals the gap between the actual and the potential activity intensity at various locations. In essence, the FAMI model provides us with the estimation of the current activity intensity potential given by the movement coming from activities themselves and as a direct effect of urban form. This potential is often not fully realized and is being restrained by a multitude of external factors such as public transport, motorized traffic, or land use regulations, just to mention a few. By quantifying the gap between the potential and the actual activity intensity, we can assess the neighborhood's unutilized capacity to accommodate activities, explore its causes, and devise actions to alleviate it (Figure 85).

Figure 85. Section of the residual map for the predicted intensity of gastronomy with residuals expressed in the form of an unutilized floor area. (Blue = negative residuals, Red = positive residuals, Black = prediction on target. The line thickness expresses the amplitude of the residual).

Finally, we tested how spatial interaction between the activities themselves affects their allocation pattern. As suggested by urban economists, part of the reason why activities can be found at a specific location has to do with the activities at neighboring locations. In other words, when activities interact with each other, they create positive or negative spillover effects, which in turn, attract or repel other activities. We found such interaction to be significant in the case of gastronomy and work activities. Gastronomy in Weimar displays strong negative autocorrelation with 49% of the potential given by pedestrian movement being reduced by neighboring activities. In other words, if a person opens a restaurant, the chance that we find another restaurant on a neighboring street is reduced by 49%, given that both streets have the same amount of pedestrians passing through. We must note that the spatial autocorrelation is highly sensitive to the areal unit of analysis. This means that our findings do not reject the possibility that gastronomy clusters within the street. However, it suggests that on the street level, repelling forces are at work. In the case of the work activities, we found the spatial interaction working in the opposite direction. With the positive autocorrelation of 7%, we conclude that work activities tend to cluster on the spatial resolution of individual streets.

The consequence of spatial autocorrelation is similar to the omitted variable bias discussed before. If present and ignored, the prediction model and its estimates are biased and inconsistent (Anselin, 2013). Since the spatial autocorrelation depends on the activity type, it seems to be necessary to always test for its presence before estimating the statistical model.

6.2 Relevance and Future Work

We started this research by the simple realization that different disciplines offer different explanations of the same phenomena. As such, this is relatively unsurprising, as different approaches are, by definition, expected to take different views. It might also come as little surprise that our empirical study found all three disciplines, CUM, TP and UE being correct in their core assumptions. We found that activities and urban form affect movement, that activities affect themselves, and are affected by urban form. So far, we could conclude that the existing theoretical models are correct and compatible with each other. However, what we consider as the crucial contribution of this study is the finding that the above-mentioned theories, propositions, and models are valid if and only if they are considered simultaneously.

In short, we found each of the individual approaches to be biased if applied in isolation. It is important to realize that a biased model is not the same as a less precise, efficient, or inconsistent model. Bias means that the model and its outcomes are structurally wrong and misleading. Thus, if there is one thing to take away from this study, it is that specialists need to look beyond their field of expertise. Or, as Piketty puts it, "they must set aside their contempt for other disciplines and their absurd claim to greater scientific legitimacy, despite the fact that they know almost nothing about anything" (Piketty, 2017, p41). Here we do not question the knowledge brought by transportation planners, urban morphologists, and urban economists. We argue, however, that it might be of little use on its own.

The need to simultaneously consider the effect of activities, movement, and urban form implies that conventional analytical prediction models are not appropriate when only information on the urban form is available. These findings present significant difficulty in urban planning since the allocation of activities might be unknown at its early stages. We argue that in such a case, alternative approaches based on numerical simulation might offer a promising alternative. The analytical model based on existing urban environments can be used to estimate the parameters of the simulation model (e.g., who is interacting with whom and how). The simulation model can be, in turn, used for future predictions. The prototype of such a form-activities-movement simulation model developed by the authors (Bielik et al., 2019) demonstrates the feasibility of the approach as well as its challenges. Besides the computational complexity of the simulation, the main challenge is the absence of empirical data, which could be used to set the model parameters. In other words, we need to know how movement affects activities, and activities affect movement, and activities affect

themselves before simulating their interactions. This study provides the first sample of such interaction parameters, which still needs to be extended and confirmed by future studies.

We argue that such analytical and predictive models are the critical components in the effort of promoting pedestrian movement, increasing the accessibility to services, and reducing the carbon footprint of our cities. They give us tools to identify unutilized potential as well as predict the impact of future decisions.

Finally, we must mention the unique role of pedestrian movement in this study and its broader consequences. Back in 2015, when this research took off, we knew that the future of the movement in cities is shifting from cars back to pedestrians. The direction was clear at least since 1961 after Jane Jacobs published her "The Death and Life of Great American Cities." Almost 60 years later, the negative effect of car-oriented planning on life quality, health, social relationships, and the environment is well known and without dispute. However, when the COVID-19 pandemic hit the world in 2020, it became painfully clear that pedestrian friendly cities are not an option but a fundamental need. The ability to walk out of home and reach essential services turned from academic discussion into a real-life necessity. Cities and neighborhoods providing such walkable environments proved their resilience and grew into a critical instrument and healing islands in the fight against the "invisible enemy." With this in mind, we hope to contribute with this study to our understanding of how long-term planning decisions affect the walkability of urban environments and to help in the ongoing process of their transformation.

Appendix

Appendix 1 Spatial Point Process

In geo-statistics, the process driving the allocation of discrete events such as the allocation of activities is called a spatial point process (SPP). The SPP is a stochastic process based on a spatial probability field (Figure 86a) controlling the intensity of the process (Figure 86b). It is important to realize that due to the stochastic nature of SPP, the same process with the same probability field produces different point patterns (Figure 86c). Thus, if we want to understand the consequences of any given SPP, we look at the aggregate of the different realizations – the potential field (Figure 86d). Based on the type of the process, the aggregated potential field, and the spatial probability field can differ.

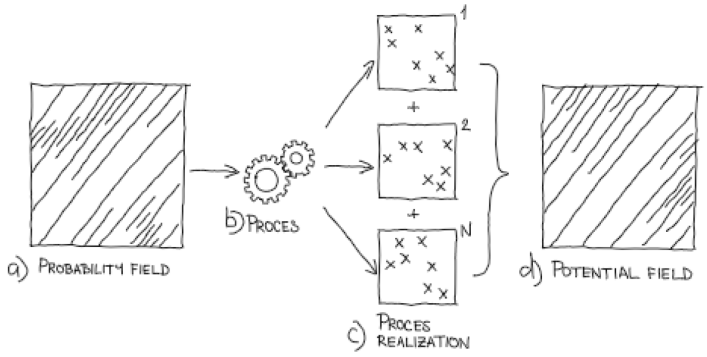


Figure 86. Spatial point pattern process. Scheme showing the difference between process intensity, process realization, and potential field.

However, in the special case, when the spatial probability field of the SPP is constant throughout the space (e.g., exogenous movement), the resulting potential field will also be constant (Figure 87).

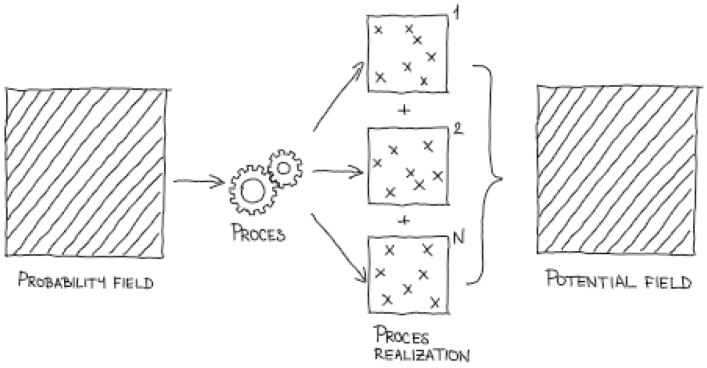


Figure 87. Spatial point pattern process driven by homogenous probability field as in the case of exogenous movement.

Appendix 2 Omitted Variable Bias

Omitted variable bias is a product of model misspecification. It occurs when a regression model specification ignores relevant explanatory variables, which are known as confounding variables (Figure 88). As a result, the model attributes the effects of the omitted variables to the cofounding variables that are in the model.

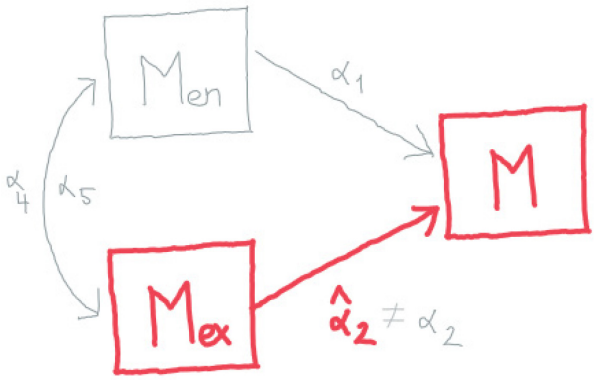


Figure 88. Schematic sketch of conditions required for the omitted variable bias. It depicts the relationships between the explanatory variables X and Z and the dependent variable Y. Considered variables are depicted in red, ignored variables are shown in grey. The estimated coefficient $\hat{\alpha}_2$ is different from the true coefficient α_2.

For example, we can imagine the following regression model specification with one dependent variable y and two explanatory variables x and z and the error term μ_1:

$$y = \alpha_0 + \alpha_1 x + \alpha_2 z + \mu_1 \tag{31}$$

Additionally, x and z are correlated, which can be expressed as:

$$z = \alpha_3 + \alpha_4 x + \mu_2 \tag{32}$$

When the variable z is ignored, the actual model specification which is being estimated is the result of the substitution of the second equation into the first one:

$$y = (\alpha_0 + \alpha_2\alpha_3) + (\alpha_1 + \alpha_2\alpha_4)x + (\alpha_2\mu_2 + \mu_1) \tag{33}$$

The problem is that when we estimate this model, the coefficient describing the impact of x on y will also include the coefficient describing the impact of z on y. Without going into mathematical details, we can say that the estimate is biased by picking up the variance of the omitted variable. This bias can result into a) overestimating the strength of an effect, b) underestimating the strength of an effect, c) changing the sign of an effect, or d) masking an effect that actually exists.

The actual type of bias depends on the relationship between the omitted and included variables and can be categorized as follows:

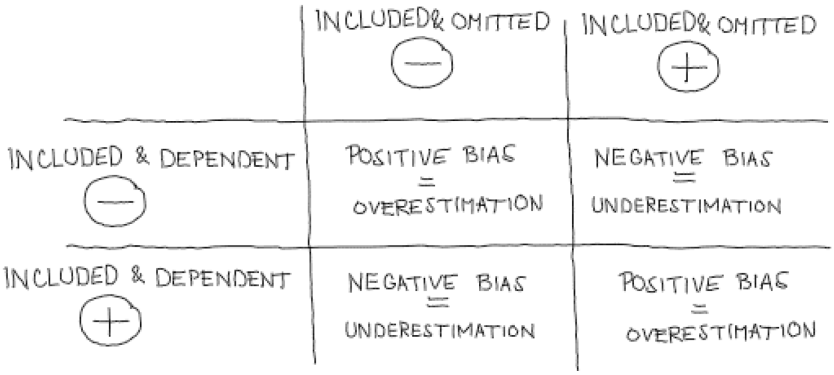


Figure 89. Matrix of possible types of omitted variable bias based on the direction of the relationship between omitted, included, and dependent variables.

Appendix 3 Simultaneity Bias

The major difficulty in the estimation of the activity-movement interaction model is the simultaneity of relationships between the explanatory and dependent variables. On the one hand, we expect the distribution of the movement to influence the activity allocation while, on the other hand, the allocation of activities to determine how people move. Moreover, we expect the activities to influence themself (Figure 90).

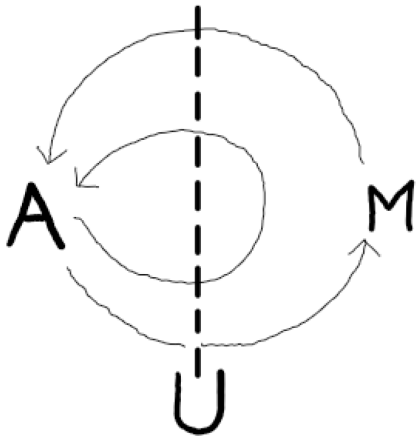


Figure 90. Simultaneous relationships between activities (A) and movement(M) through the urban form (U).

In general, simultaneity means that not only the explanatory variable is affecting the dependent variable, but the relationship goes simultaneously both ways. As a result, both variables can be considered as dependent, which causes that there is not enough information to tell apart the effect of x on y from the effect of y on x. Such a model violates basic assumptions formulated in the Gauss-Markov theorem specifying under which the Ordinary Leas Square (OLS) the best linear unbiased estimator (BLUE).

In our case, there are two sources of simultaneity which introduce bias in the estimated regression equation. On the one hand, we assume simultaneous spatial interaction between activities (i.e., autocorrelation). On the other hand, there is the simultaneous activity-movement interaction. In the case of autocorrelation, the simultaneity is above obvious. We literally have the same variable on both sides of the equation. Correcting this bias requires the introduction of a different regression model (Spatial Autoregressive Linear Model - SARLM), using a different estimator (Maximum likelihood, or General method of moments) and additional variables (Weight matrix). In other words, it is possible but requires additional assumptions that must be tested.

In the case of the endogenous movement, the argument is trickier. At first glance, it seems obvious that simultaneity takes place as we model the effect of movement on activities while expecting the activities to affect the movement. As a result, to avoid the simultaneity bias

when modeling the effect of movement on activity levels, we need to correct for the simultaneous effect of activity on movement and vice versa. However, we argue that this feedback does not cause bias as the effect of activity on movement is highly non-linear. As discussed in the Methods and Data section, how activity levels on a given location materialize in the actual movement depends on additional parameters such as a) street network configuration and b) activity levels at other locations. As a result, streets which offer only a few activities can still be highly frequented. Such movement is not originating nor ending at a given street, but it is passing through. In order for the simultaneity to cause bias when estimating the linear effect of movement on activity, it is required that also activity allocation linearly affect movement. In such a case, the additional instrumental variables are needed to separate both simultaneous effects from each other. However, in our case, these effects are not both linear, and thus no additional correction is required.

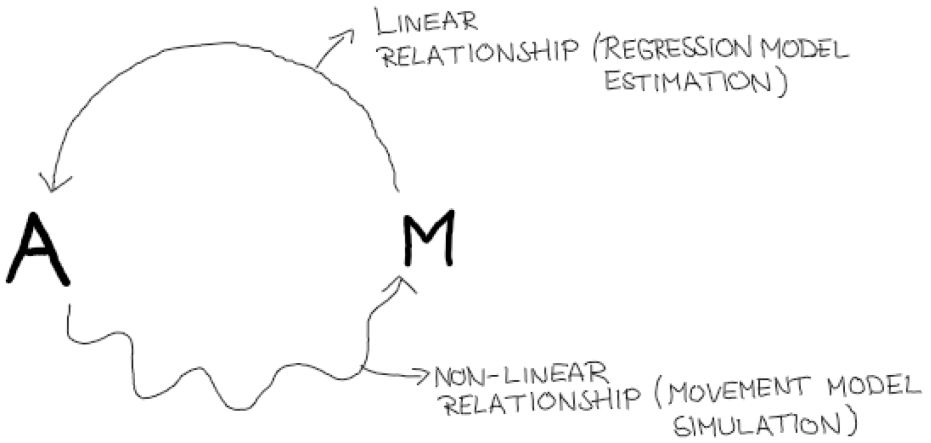


Figure 91. Illustrating the different character of the relationships of the activity-movement interaction model.

Appendix 4 Building Geometry Data

The data is provided by the "Landesamt für Bodenmanagement und Geoinformation" and was accessed on Mai 2018 via "Offene Geodaten Thuringen." The total count of the building objects for the administrative area of Weimar is 34 871, with a total footprint of 3 123 879 m^2 and a floor area of 7 816 130 m^2.

Buildings are represented as footprint extrusions, with one building consisting of one or several such extrusions (e.g., the main house, guesthouse). Unfortunately, there is no information about which extrusion objects belong to one same building. Each extrusion is one entry in the database with the following attributes: Height, Footprint, Administrative area, and Function.

Appendix 5 Activity Allocation Data

In this study, only activities that take place in buildings are considered. The data on their allocation come from two data sources: a) the governmental open geodata on building geometry and function - geoportal-th.de and b) non-governmental collaborative geodatabase – openstreetmap.org. These data sources are first filtered by activities relevant as movement attractor identified by the MiD2017 study (Appendix 7). Next, both data sets are examined on the duplicates and finally combined into one. The use of multiple data sources was necessary as each of these sources uses a different activity categorization scheme and is better at capturing some of the activities than the others.

The governmental dataset categorizes the building functions based on the functional standard defined in the ALKIS-Objektartenkatalog version 6.0. The ALKIS scheme is less detailed than the OpenStreetMap categorization scheme, especially when it comes to commercial activities. It does not differentiate between different types of commercial activities such as errands, bars, restaurants, and other commercial services. In total, only 49% of the building objects have defined function, which can be attributed mainly to the fact that one building is often divided into multiple objects with only the main object containing all functional attributes. For the building objects with function tag, we found 22 functional categories relevant as pedestrian movement attractor (see Appendix 7) and two tags representing the origins of pedestrian movement. We group them into broad categories following the MiD2017 travel activity scheme. These are accommodation, administrative, healthcare, education, and work activities (see Table 12). From the governmental data set, we were not able to differentiate between daily shopping and gastronomy.

Table 12. Categorization of functional tags for Geoportal-th building objects in the study area.

Destinations of movement					Origins of movement
Gastronomy & Shopping	Education	Administrative	Health	Work	Accommodation
Business and commerce	Vocational school	Administration	Hospital	Business and commerce	Residential building
Trade and services	University building	Public buildings	Healthcare	Trade and services	Mixed-use with living
	General school	Post Office		Vocational school	
		Town hall		University building	
				General school	
				Administration	
				Public buildings	
				Post Office	
				Town hall	
				Hospital	
				Healthcare building	
2753	247	232	18	3250	16156

The openstreetmap.org (OSM) data set is collaboratively mapped by a large user community [33] and offers a rich categorization scheme for building functions and associated activities. In contrast to the governmental data set and the MiD2017 activity scheme, the OSM offers the highest level of detail. As a result, we aggregate several OSM categories (defined by key, value tags) into six movement relevant activities (Table 13).

[33] 6,450,058 contributors worldwide source: "OpenStreetMap Statistics", OpenStreetMap Foundation. Retrieved 11 May 2020.

Table 13. Categorization of OSM activity tags in the study area.

Destinations of movement

Gastronomy		Shopping		Education		Administrative		Health		Work	
Key	Value	Ke	Value	Key	Value	Key	Value	Key	Value	Key	Value
amenity:	restaurant	shop:	alcohol	amenity	college	amenity	bank	amenity	clinic	building	commercial
amenity:	cafe	shop:	bakery	amenity	language_sch	amenity	post_office	amenity	dentist	building	industrial
amenity:	bar	shop:	beverages	amenity	music_school	amenity	post_depo	amenity	doctors	building	kiosk
amenity:	biergarten	shop:	butcher	amenity	school	amenity	post_box	amenity	hospital	building	office
amenity:	Fast_food	shop:	cheese	amenity	university	amenity	townhall	amenity	pharma	building	retail
amenity:	food_cour	shop:	chocolate	building	college	amenity	atm	building	hospital	building	supermarket
amenity:	ice_cream	shop:	confectione	building	kindergarten					building	warehouse
amenity:	pub	shop:	convenienc	building	language_sch					building	bakehouse
		shop:	frozen_foo	building	library					building	civic
		shop:	greengroce	building	music_school					building	government
		shop:	heaalth_fo	building	school					building	hospital
		shop:	pastry	building	university					building	kindergarten
		shop:	tea							building	public
		shop:	water							building	school
		shop:	departmen							building	train_station
		shop:	kiosk							building	transportation
		shop:	supermark							building	university
										office:	accountant
										office:	advertising_agency
										office:	architect
										office:	association
										office:	bail_bond_agent
										office:	charity
										office:	company
										office:	consulting
										office:	coworking
										office:	diplomatic
										office:	educational_institu
										office:	employment_agenc
										office:	energy_supplier
										office:	engineer
										office:	estate_agent
										office:	financial
										office:	forestry
										office:	foundation
										office:	geodesist
										office:	government
										office:	guide
										office:	insurance
										office:	it
										office:	lawyer
										office:	logistics
										office:	moving_company
										office:	newspaperngo
										office:	notary
										office:	property_managem
										office:	religion
										office:	research
										office:	surveyor
										office:	tax
										office:	tax_advisor
										office:	telecommunication
186		100		138		109		71		1488	

By further examination of the OSM data set, we conclude that it outperforms the governmental data on most of the public activity types; however, it is relatively coarse when it comes to private activities such as accommodation. On the contrary, the accommodation is well-mapped in the governmental data, which is the reason why we combination of both sources is considered useful.

To combine the two data sets, we use the building geometry acquired from the geoportal-th as a basis and project the OSM activities as point data over it. As next, we assign all OSM points inside a building boundary to the respective buildings and combine it with the existing database entries. In general, there are three cases we consider:

a) There is no existing database entry for a given building. In this case, the OSM activity tag is imported into the database
b) There is an existing database entry for a given building, and it corresponds with the OSM activity tag. In this case, nothing happens.
c) There is an existing database entry for a given building, and it is different from the OSM activity tag. In this case, a new activity is added. Such a building contains multiple activities at the same time (e.g., accommodation and shopping).

Appendix 6 Activity Types - Dimension Reduction

In the following, we investigate if the set of six travel activity types[34] cannot be reduced to simplify the activity-movement interaction model. We define each street segment by vector of six activities intensities associated to the particular street. The question we investigate here is if this vector cannot be reduced by maintaining its information value. We conduct exploratory or "unrestricted" factor analysis (FA) to extract reduced set latent variables (i.e., underlying concepts) explaining the variance in the measured variables.

Determining the Number of Factors (Latent Variables)

There are multiple approaches to identify the optimal number of factors, such as the Scree plot, Kaiser's rule, and Horn's parallel analysis. They all calculate factor analysis with the full set of factors and then look at the proportion of variance explained by each factor. The full set is than used to determine the number of factors which is meaningful to describe the data. This means that adding more factors will produce only marginal improvement of the explained variance and might compromise their interpretability.

In the Scree plot (Figure 92), we look for elbows – spots beyond which the curve stays flat. The flat curve represents only a marginal contribution to the explained variance. We identify this spot after the second factor. Kaiser's rule suggests taking only factors with an eigenvalue higher than one. The eigenvalue expresses the variance explained by the factor and suggests that only the first factor should be considered (see the horizontal line in Figure 92). Finally, Horn's parallel analysis compares eigenvalues generated by the data with eigenvalues generated from the Monte-Carlo simulated matrix coming from random data of the same size. We look for the point of infliction of these two lines (i.e., where the lines are crossing) and accept only factors above this eigenvalue. In our case, this results in three factors.

[34] One type for origin - Accommodation and six types for destination of pedestrian movement – Administration, Education, Gastronomy, Health, Shopping, Work.

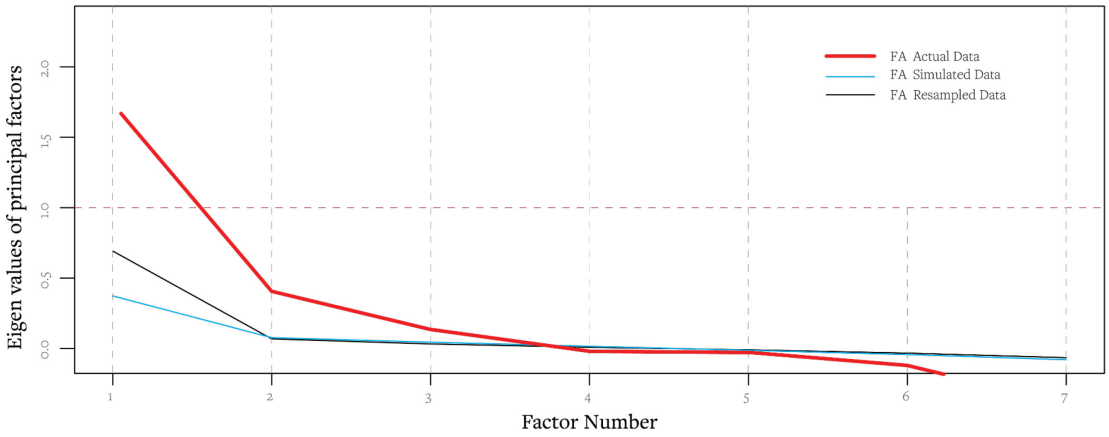


Figure 92. Parallel analysis Scree plot.

We conclude that the different methods for determining the optimal number of factors suggest between 1 to 3 factors. We decide for the higher bound of the range – 3 factors to maximize the explained variance by the factor variables.

Interpretation of Factors

We conduct the factor analysis with a reduced set of three factors. Based on the previous empirical analysis and theory, we cannot assume complete orthogonality (i.e., independence) of the activity factors. In other words, we do expect the activities to be correlated and not being completely independent of each other. For this reason, we adopt oblique matrix rotation – Oblimin resulting in the factor loadings presented in Table 14.

Table 14. Standardized loadings (pattern matrix) based upon correlation matrix (loading cut-off = 0.3)

	Factor 1	Factor 2	Factor 3	Communality
Accommodation	-0.03	0.07	**0.46**	0.23
Work	0	**1**	0.01	1
Education	-0.06	**0.4**	-0.09	0.15
Shopping	**0.71**	-0.02	0.11	0.55
Administrative	**0.58**	0.04	-0.04	0.34
Healthcare	**0.77**	0	-0.1	0.56
Gastronomy	0.41	0.04	**0.3**	0.34

The factor loading matrix can be described as structurally simple, which means that each of the original variables loads only on one factor at a time. This allows us to interpret the Factor 1 latent variable behind the Shopping, administrative, healthcare, and gastronomy travel activities. Factor 2 is explaining the variation of work and educational activities, and factor 3 stands for the accommodation.

Efficiency of Factors

When examining the communalities – the extent to which the selected factors account for the variance of measured variables, we found that in 4 out of 7 cases, its value lies below 0.5. This means that less than half of the variation in the measured variables is captured by the factors. This is reflected in the cumulative variance explained by the three factors together, also being below the threshold of 0.5 (Table 15).

Table 15. Variance explained by each of the three factors

	Factor 1	Factor 2	Factor 3
Cumulative variance	0.23	0.4	0.45

Summary

The factor analysis does reveal a reduced set of latent variables. These are, however, able to capture only less than half of the variation of the measured variables. In other words, more than half of the information gets lost on the way. The ability of latent factors to capture the variance of measured variables does not improve by increasing the number of factors and stays below 0.5. Therefore, we conclude that the factor analysis was not able to identify latent variables, which at the same time a) reduce the complexity, b) maintain interpretability, and c) keep the information value of the measured data on activity distribution. For this reason, we keep the original seven activity categories for the purpose of this study.

Appendix 7 Movement Data

The data on pedestrian movement comes from three different sources, each capturing different aspects (e.g., trip length, trip frequency) and resolution (e.g., city, neighborhood, street) of pedestrian movement. The combination of these data sources allows us to estimate movement model parameters and the relationships between the exogenous and endogenous movement components generated by each activity type. Moreover, this data is used to quantify the impact of various model restrictions (e.g., limiting the number of travel activities) on the expected accuracy and usefulness of the activity-movement interaction model.

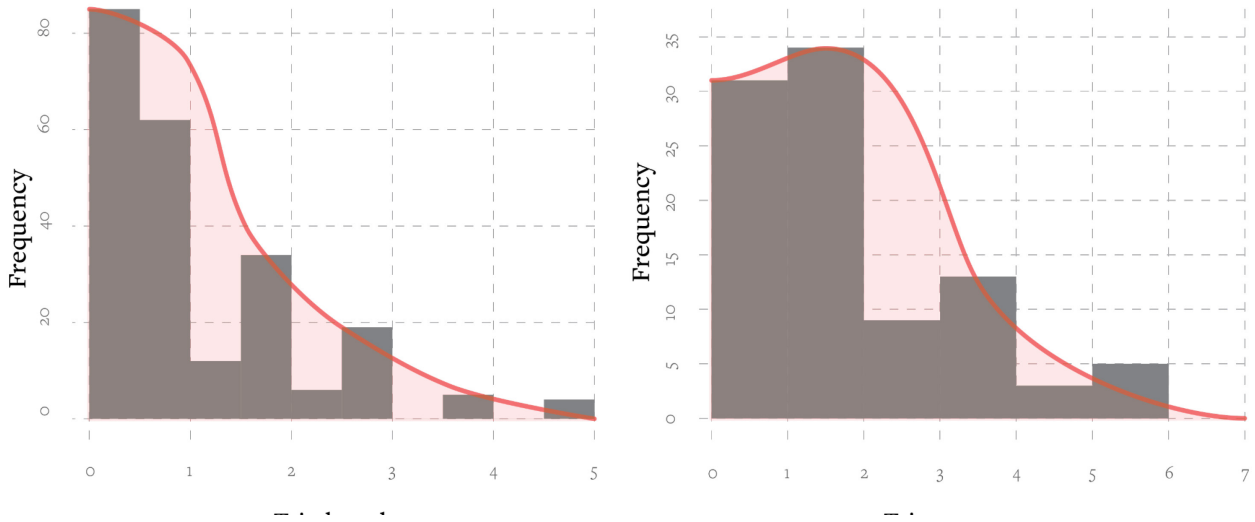
Mobilität in Deutschland 2017 (Mobility in Germany 2017)

The nationwide mobility study "Mobilität in Deutschland 2017" (MiD2017) captures the movement flows on three levels of spatial resolution - country, regional and local. It contains information about the activity at the origin and destination of each trip as well as its time, length, and mode of transport. For the purpose of this research, we use the MiD2017 data at its finest resolution – 500x500m grid defined by "Bundesamt für Kartographie und Geodäsie."[35]

The selected data grid cells for Weimar contain a single-day travel diary of 180 individuals resulting in 654 individual journeys. From these journeys, we found 248 pedestrian trips taken by 98 individuals. The mean pedestrian trip length is 1.20, with a standard deviation of 1.04. The distribution of pedestrian trip lengths is unimodal and strongly right-skewed with gradual decay of trip frequency as the trip length increases (Figure 93a). The average daily pedestrian trip count is 2.53 trips per day, with a standard deviation of 0.76 trips[36]. The distribution has similar attributes as in the case of the trip frequency with strong right skew as most people take no more than two trips a day (Figure 93b).

[35] Accessed from https://gdz.bkg.bund.de/ in September 2019

[36] Only individuals who walked were considered.



a) b)

Figure 93. Histogram and density plot of a) pedestrian trip length and b) pedestrian trip count per day in Weimar (MiD2017).

To simplify the interaction model, we analyze the contribution of each travel activity to overall movement and identify a set of activities that can be left out with only a marginal impact on the model coverage. In total, we found 25 activities (Table 16) from which the interaction model cannot represent 13 as they are either a) not bound to any location, with the travel being the purpose on its own (e.g., walking dog, jogging) or b) the location of the travel activity cannot be identified (e.g., work trip, private business). This set of 13 activities accounts for 40,31% of all trips and cannot be represented by the interaction model as it requires information on both, the origin and destination of movement.

From the remaining 12 travel activities, six attract below 3% of all trips. To simplify the model in terms of the number of variables and interactions, which must be estimated, we restrict it to only six core travel activities – administrative, education, gastronomy, healthcare, shopping, and work activities, which together account for 45% of total trips.

Table 16. MiD2017 Travel activities with their respective contribution to the total trip count in the study area.

Travel purpose (Original MiD2017 description)	Travel purpose (Translation)	Trip count (%)	Considered in the movement model
Arbeit	**Work**	**6.00%**	**YES**
Dienstlich	Business	3.00%	NO (Unknown location)
Ausbildung	**Education**	**8.00%**	**YES**
Täglicher Bedarf	**Shopping daily goods**	**16.96%**	**YES**
Dienstleistungen (Friseur, Schuster etc.)	Services (e.g., Hairdresser)	0.72%	NO (Marginal contribution)
Arztbesuch, andere medizinische Dienstleistungen	**Medical visit and services**	**5.82%**	**YES**
Behörde, Bank, Post, Geldautomat	**Authority, bank, post office, ATM**	**4.36%**	**YES**
Private Erledigung für andere Person (unentgeltlich)	Handling of private issues for others	1.45%	NO (Unknown location)
Sonstiger Erledigungszweck	Other issues	4.36%	NO (Unknown location)
Besuch/Treffen von Freunden, Verwandten, Bekannten	Visit/meeting with friends, relatives, acquaintances	5.90%	NO (Unknown location)
Besuch kultureller Einrichtung (z.B. Kino, Theater, Museum)	Visiting cultural institutions (e.g., cinema, theater, museum)	0.66%	NO (Marginal contribution)
Besuch einer Veranstaltung (z.B. Fußballspiel, Markt, Popkonzert)	Attending an event (e.g., football match, market, pop concert)	1.31%	NO (Unknown location)
Sport (selbst aktiv), Sportverein (z.B. Fußball, Tennis, Training, Wettkampf)	Sport activity (e.g., football, tennis, training, competition)	2.62%	NO (Marginal contribution)
Restaurant, Gaststätte, Mittagessen, Kneipe, Disco	**Restaurant, pub, lunch, pub, disco**	**3.93%**	**YES**
Tagesausflug, Kurzreise bis zu 3 Übernachtungen	Day trip, a short trip up to 3 nights	0.66%	NO (Unknown location)
Spaziergang, Spazierfahrt	Walk, drive	10.49%	NO (Unknown location)
Hund ausführen	Walking dog	9.84%	NO (Unknown location)
Joggen, Inlineskating etc.	Jogging, inline skating etc.	0.66%	NO (Unknown location)
Kirche, Friedhof	Church, cemetery	1.31%	NO (Marginal contribution)
Ehrenamt, Verein, politische Aktivitäten	Volunteer work, association, political activities	0.66%	NO (Unknown location)
Hobby (z.B. Musizieren)	Hobby (e.g. making music)	0.66%	NO (Unknown location)
Spielen auf der Straße etc.	Playing in the street etc.	0.66%	NO (Unknown location)
Sonstiger Freizeitzweck	Other leisure purpose	0.66%	NO (Unknown location)
Model coverage		**45.1%**	

To further simplify the movement model, we restrict the considered origins to home locations only. To evaluate the proportion of pedestrian movement covered by this restricted movement model, we calculate the proportion of pedestrian trips between home and any of the six considered travel activities. In our case, the direction of the movement (i.e., activity to home or home to activity) does not play any role and is considered perfectly symmetrical (i.e., the same shortest path is chosen in both directions).

We found that the restricted movement model accounts for 42,93% of all pedestrian trips, with a 95% confidence interval[37] between 35,68% and 56,21% (Figure 94).

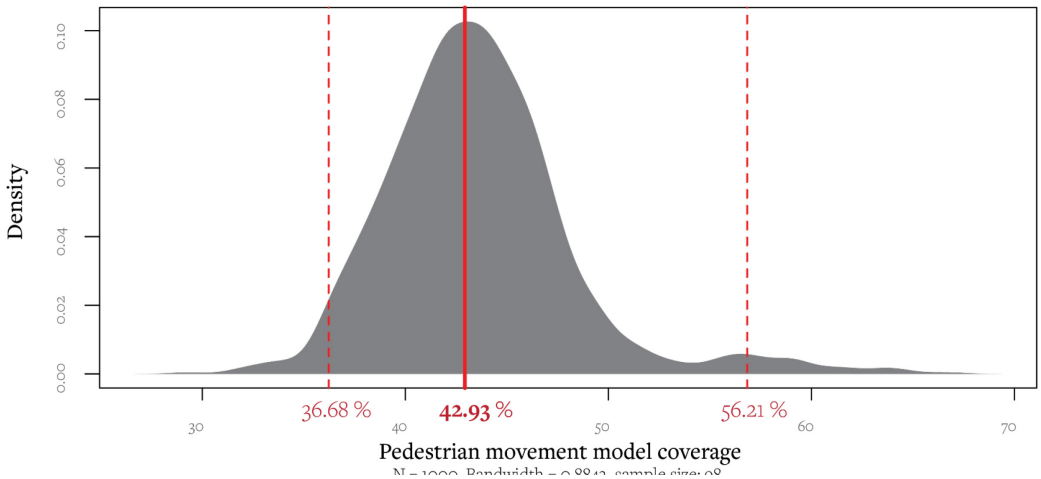


Figure 94. Mean coverage (red line) and 95% confidence interval (dashed red lines) of pedestrian movement model restricted by the six travel destination activities and accommodation as only activity on the travel origin.

The remaining 57% of trips are either trip chains or are not based on any of the selected travel activities. It must be noted that this does not necessarily mean that 57% of the variance in pedestrian movement is not captured by the model, but rather that we can guarantee that the model captures at least 43%. The explained variance can be higher if a) the captured movement coincides with the trip chain between the home, secondary activity, and primary activity or b) the distribution of any activity left out from the model follows the distribution of some of the included activity. In such a case, the resulting model will have higher explanatory power but might run the risk of omitted variable bias and parameter overestimation (see Appendix 2).

[37] We calculate the confidence interval via bootstrapping (Bradley Efron, 1979) also known as non-parametric Monte Carlo random sampling method (Stanislaw Ulam, 1940). The major difference between the two is that Monte Carlo simulates data and bootstrapping takes the data as given and just resamples it over and over. What is advantageous about bootstrapping is that no assumption about the underlying distribution or its properties is assumed. In bootstrapping it is assumed that all observations come are independent and identically distributed – this is fulfilled by the design of the MiD2017 study. The bootstrapping parameters were set to 1000 simulations with population 98 (size of the original sample) per simulation.

Shortest Path Study 2017

In 2017, we conducted an empirical study on pedestrian route choice behavior with the goal of testing the different theoretical assumptions and inform the pedestrian movement simulation model. Participants of this study (N=50) were students with a high degree of familiarity with the case study area (longer than one year). All participants were asked to keep a detailed travel diary recording each journey, travel purpose, time of departure, and transportation mode (Figure 95). Our methodology was based on the MiD2017 study capturing multi-purpose journeys as individual routes with their respective travel purpose. In total, we collected 529 routes for four transportation modes (Walking, Cycling, Bus, Car) and seven travel purpose categories (Accommodation, Work, Education, Shopping, Administrative, Healthcare, Gastronomy) matching the activity types adopted in this research (see Appendix 7). All journeys were digitalized and projected over the street network data from Geoportal-th (Appendix 8).

Route Choice Study | ID 51 CUA | Weimar 04 - 2017

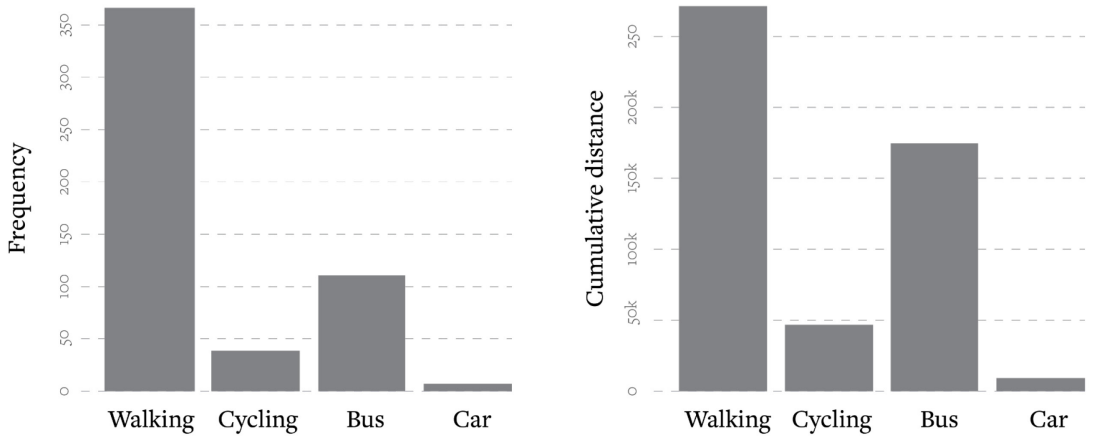
Date: 11.04.2017 | **Hour**: 16 : 30 | **Travel purpose:** Home, Work, Education, Shopping, Eat&Drink, Culture, Sport, Leisure, _____

Legend:
Origin **O** | Destination **D** | Walking **W** | Cycling **C** | Bus **B** | Car **A**

Figure 95. Exemplary route choice 2017 study material.

Figure 96. Set of 203 unique pedestrian paths as captured by the Path selection 2017 study. Paths are laid on top of each other, with the color intensity indicating multiple paths sharing the same street segment.

Overall, we found that most trips were pedestrian (71%), followed by public transport (21%), cycling (7%) and with only 1% of car trips (Figure 97). If we translate the results in the traveled distance, we found that 54% can be attributed to walking, 35% to public transport, 9% to cycling, and 2% to car travel.



a) b)

Figure 97. Bar plot showing the distribution of a) trip counts and b) travel distance by individual travel modes.

Pedestrian Counting Study 2016

To estimate the portion of movement caused by endogenous and exogenous movement respectively, we collected in March 2016 empirical data on pedestrian *Through-Frequency* (i.e., number of pedestrian passing through) at 100 locations distributed across the study area (see Figure 98).

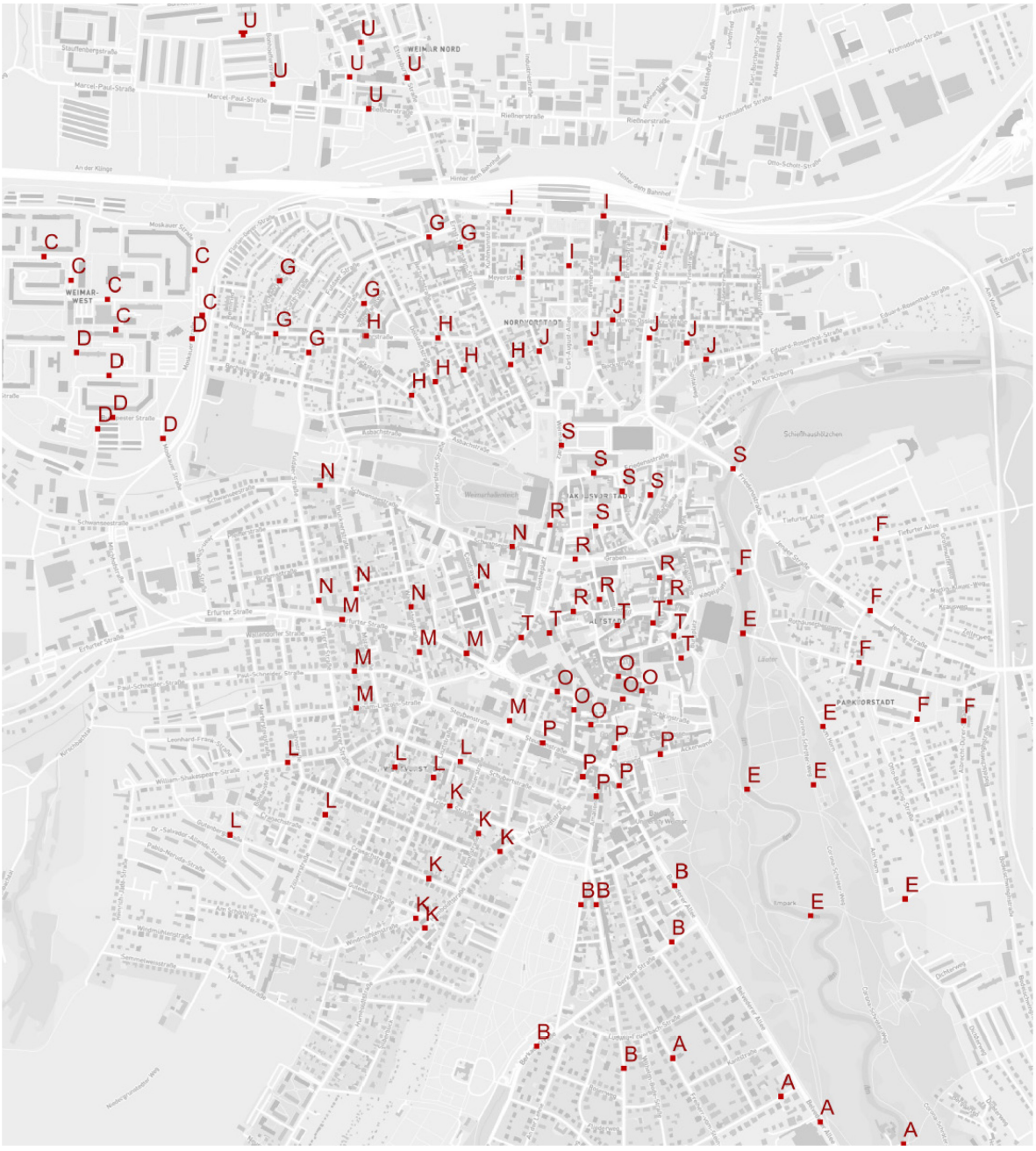


Figure 98. Pedestrian counting 2016 study locations. Map is showing the 100 counting locations considered in the study.

A 1 | Counting sheets Weimar 03 - 2016

Working day (Day/Month): ...../.....

	Start time (8:00-10:00):		Start time (12:00-14:00):		Start time (16:00-18:00):	
	Pedestrian:	Cyclist:	Pedestrian:	Cyclist:	Pedestrian:	Cyclist:
RAW DATA						
SUM						
NOTE						

Working day (Day/Month): ...../.....

	Start time (8:00-10:00):		Start time (12:00-14:00):		Start time (16:00-18:00):	
	Pedestrian:	Cyclist:	Pedestrian:	Cyclist:	Pedestrian:	Cyclist:
RAW DATA						
SUM						
NOTE						

Figure 99. Exemplary pedestrian counting 2016 study material for marking the counts at different daytimes and weekdays.

The pedestrian flow at each location was measured for 15 minutes time period on three days (two working days and one weekend day) and three different time slots (8-10, 12-14, 16-18) [38]. As a result, we collected nine measurements for each location, which allow us to analyze the variation in pedestrian flow throughout the weekdays and daytime. To select an appropriate time unit in which the pedestrian flow should be represented, we analyze the variation in the pattern as well as the variation in the amplitude across all nine measurements.

The variation in the pattern means that during different measurements, the distribution of movement flow changes. It would suggest that the movement simulation model must incorporate the changes in travel supply and demand patterns over time. Variation in the amplitude suggests that a more stable temporal unit of the measurement must be selected.

We explore the variation in the movement pattern via correlation matrix comparing the linear relationship between different daytimes and weekdays. If the pattern remains stable (i.e., variation is low), we expect values close to one. On the contrary, values close or equal to zero suggest no linear relationship between the two movement patterns. The results of Pearson's correlation matrix (Table 17) reveal a strong and significant relationship between all nine patterns. The correlation coefficient ranging from 0.6 to 0.96 and being centered at 0.8 suggests that all measurements can be represented via a single pattern.

Table 17. Pearson's correlation matrix showing the correlation coefficient R for nine pedestrian counting sessions. Significant correlations are marked with * (p-value ≤ .05)

	Weekend 8-10 am	Weekend 12-2 pm	Weekend 4-6 pm	WD. 8-10 am	WD. 12-2 pm	WD. 4-6 pm	WD2. 8-10 am	WD2. 12-2 pm
Weekend 12-2 pm	0.728*							
Weekend 4-6 pm	0.674*	0.956*						
Working Day 8-10 am	0.689*	0.845*	0.88*					
Working Day 12-2 pm	0.646*	0.873*	0.908*	0.864*				
Working Day 4-6 pm	0.688*	0.887*	0.922*	0.903*	0.928*			
Working Day2 8-10 am	0.659*	0.714*	0.726*	0.774*	0.67*	0.757*		
Working Day2 12-2 pm	0.675*	0.839*	0.892*	0.85*	0.959*	0.911*	0.692*	
Working Day2 4-6 pm	0.602*	0.746*	0.866*	0.835*	0.868*	0.861*	0.688*	0.911*

[38] The counting was conducted by counting for 15 minutes at one location and then changing to the next location. Consequently, for each counting we know only that it was conducted in the two-hour time slot but not the exact time.

This is confirmed by the principal component analysis showing that the first principal component explains over 90% of the total variance in the data (see Table 18 and Figure 100).

Table 18. Principal component analysis of measurements taken at the same location but at different daytime and weekday.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	160.71	36.94	21.97	17.78	15.65	11.95	10.84	9.93	6.12
Proportion of Variance	0.90	0.04	0.01	0.01	0.00	0.00	0.00	0.00	0.00
Cumulative Proportion	0.90	0.94	0.96	0.97	0.98	0.99	0.99	0.99	1

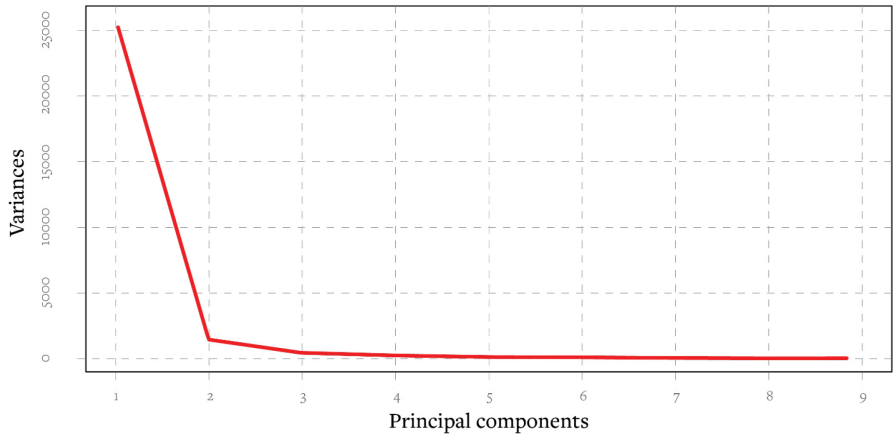


Figure 100. Explained variance by each principal component. The elbow plot suggests that only the first principal component is required to represent the variance in the data.

We conclude that all nine measurements depict the same underlying pattern, and thus the variation in movement pattern over daytime or weekdays does not present any conceptual difficulty in the movement simulation model.

From here, we move to the analysis of change in the amplitude and examine how a) different daytime and b) different weekdays affect the total amount of pedestrian traffic. For this purpose, we run a series of F-tests (one-way analysis of variance - ANOVA) comparing the between-group and withing group variation. This can be best understood when looking at the changes in mean, together with its 95% confidence intervals (CI) (Figure 101). We consider that time or day of the measurement has a significant effect on the total amount of pedestrian traffic if the difference between the weekdays and daytimes is larger than the variation within each measurement (represented by the CI).

When testing the effect of daytime, we found the pedestrian flow in the morning significantly lower than during the midday and afternoon (Figure 101, bottom row). This observation

was confirmed by all three F-test's with f-value = 3.57 and p-value = 0.029 for working day 1, f-value = 7.05 and p-value = 0.001 for working day 2 and f-value = 7.39 and p-value < 0.001 for working day 1. We conclude that the effect of daytime was strongest during the weekend, however relatively consistent across all three days. Furthermore, we conclude that we found only a marginal difference between the midday and afternoon traffic frequency.

The relatively weak effect of the weekday on the amplitude of movement can be seen in Figure 101 (top row). We conclude that the F-test did not find any significant effect of weekday when it comes to movement during the midday (f-value = 0.62 and p-value = 0.536) and afternoon (f-value = 0.405 and p-value = 0.667). However, in the morning, we measured significantly lower pedestrian flow at the weekend when compared to the weekdays (f-value = 5.98 and p-value = 0.002).

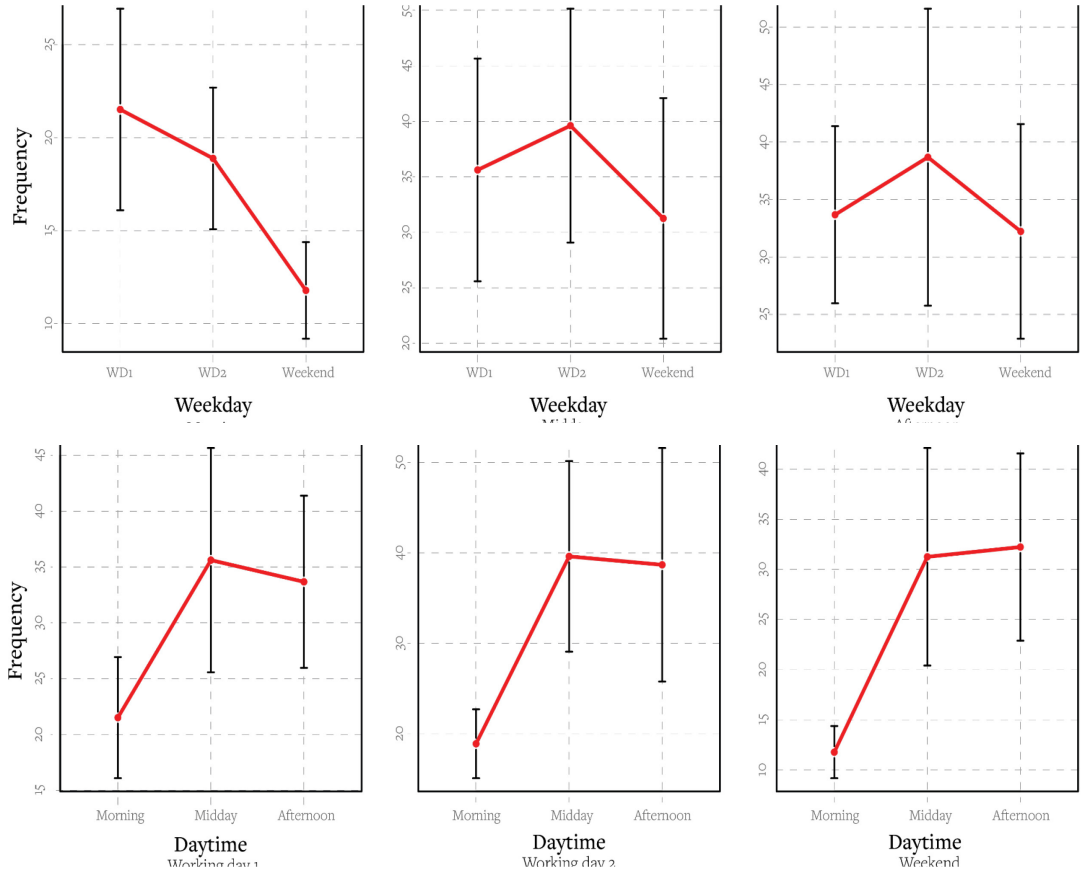


Figure 101. Change in the mean pedestrian flow across the weekday and daytime. The blue bar marks the 95% confidence interval.

In summary, we found that while the movement pattern is stable across time, the amplitude tends to change. It varies during the daytime but is relatively stable across weekdays. The exception is the amplitude of movement during the morning hours, which seems to be different between the workdays and the weekend.

Appendix 8 Street Network Geometry Data

The data is provided by the "Landesamt für Bodenmanagement und Geoinformation" and was accessed on Mai 2018 via the online portal "Offene Geodaten Thuringen."[39] The raw dataset was automatically converted to the segment map standard (see Chapter X.X). We applied a set of algorithms (Figure 102) to a) turn the street network into a valid graph representation, b) make the street network suitable for both metric and cognitive shortest path analysis, and c) reduce its size to speed up the analysis.

We begin with splitting all lines at their intersections so their topological relationships can be correctly represented as a graph (Figure 102). Then, we remove duplicate and parallel lines from the data set and simplify the crossing's complexity by applying a snapping algorithm (snapping threshold equals to 5m). This reduces the number of turns necessary to pass through the crossing and its suitable approximation of the mental representation of the navigational task.

Finally, the street network size was reduced by simplification of the network in between the crossings. Here, continuous street sections are often represented by multiple line segments which can be combined into a single line. In this process, we experimentally established a 5-degree angular deviation between two neighboring line segments as an optimal trade-off between accuracy and simplicity of the street network representation.

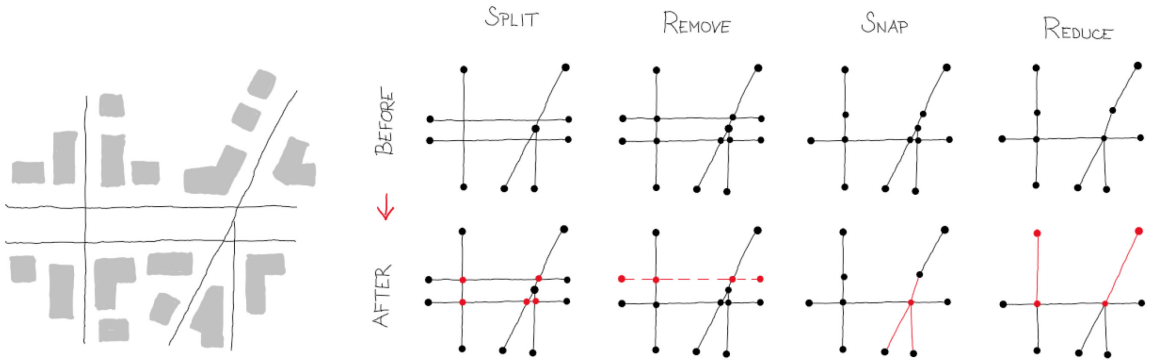


Figure 102. Sequential set of street network editing algorithms.

By applying this set of algorithms, we reduced the original data set from 12 523 to 7104 street segments. From this, 2750 are pedestrian-only (310 487 m in length), and 4354 can be used by motorized vehicles (357 156 m in length).

[39] https://www.geoportal-th.de

Appendix 9 Spatial Data Aggregation

To assure the synchronicity of all variables in the activity-movement interaction model, we map activity allocation from buildings to street segments and use these as the common analysis unit throughout this study. In the following, we discuss three methods for mapping the floor area assigned to each activity from buildings to the street network. We quantify the bias introduced by each aggregation method and select the best option (Figure 103).

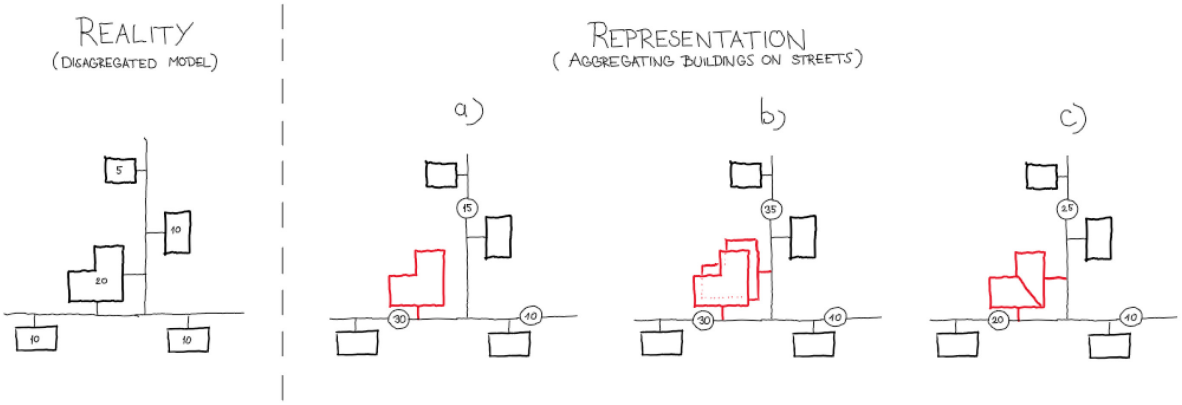


Figure 103. Three different methods for aggregating buildings on street segments. a) closest street aggregation, b) accessibility-based aggregation, c) normalized accessibility-based aggregation.

Aggregation Methods

Distance-based Aggregation

Arguably the computationally simplest and most common method with the activity weights (i.e., floor area) being assigned to closest street segments (Figure 103a). By doing this, we assume that the building is accessible only and exclusively from the closest street segment. The advantage of this method is that the sum of activity weights before and after aggregation remains constant, and it is simple to interpret the results of the aggregated model as it can be easily projected back to the individual buildings.

The disadvantage is that this method ignores the cases such as block edges or T crossings when a building is accessible from several streets at the same time. Assigning the possibility to access the building only to one of these access options introduces bias when the shortest paths are calculated as traffic leading to the building is restricted to flow through one of the multiple access street segments.

Accessibility-based Aggregation

In this approach, the activity weights are assigned to all segments from which the building is directly accessible (Figure 103b). The advantage of this approach is that the accessibility to all activities before and after aggregation remains the same. Nevertheless, the total amount of activity weight before and after aggregation might differ. Since one weight can

be assigned to multiple street segments if the building is simultaneously accessible from all of them, this weight is being multiplied. As a consequence, an activity has an unproportionally more influence on the movement flow if it is accessible from multiple streets. Additionally, in this type of model, all access points must be used equally, often even if some of them are closer to a particular origin or destination of movement than others. We conclude that in contrast to the previous *distance-based* model where the use of multiple accesses was not allowed, here, the use of all possible access streets is enforced. Consequently, both concepts violate the underlying assumption of route choice being based on the shortest path between origin and destination of movement.

Normalized Accessibility-based Aggregation

Activity weights are proportionally assigned to all segments from which the building is directly accessible (Figure 103c). In other words, if the building is accessible from two street segments, each of the two segments gets assigned half of the building weights. The advantage is that the sum of weights after aggregation equals the sum of weights before aggregation as well as the access to the activities does not change.

Nevertheless, only the portion of activity might be accessible after the aggregation procedure through the same street segments as before. In essence, activities that were accessible from multiple street segments before aggregation are divided into equal portions and exclusively assigned to the individual street segment. This creates a similar bias as in the *distance-based* aggregation model, where the whole building with its activities was exclusively assigned to one street segment.

Disaggregated "True" Model

The influence of the aggregation bias on the resulting movement model is quantified by comparing the pedestrian movement simulated by the different aggregation approaches to the movement simulated from the disaggregated model.

Empirical Study

We test the effect of the aggregation method on the street network weighting and the resulting movement simulation. As discussed previously, each aggregation method introduces a specific bias in the pedestrian movement model. We compare the movement pattern based on disaggregated – true pedestrian model (origins and destinations are individual buildings) and three different types of aggregated models. By doing so, we quantify the bias of each aggregation model and choose the least biased alternative. Due to the computational complexity of the disaggregated movement model, the test is conducted on a smaller sub-region of the study area with a size of 375 ha (2.5 km x 1.5 km). Since the goal of this experiment is to explore and quantify the differences and bias of the aggregation methods, we selected densely populated area minimizing the portion of street segments with

no buildings in their proximity. Such streets get the same zero weightings regardless of the aggregation method and, therefore, are not useful for the purpose of this study. The selected densely populated Weimar inner city area is containing 1069 street segments and 3525 buildings. For the purpose of exploring the differences and bias caused by the aggregation method, we do not differentiate between activity types but consider only the overall floor area (m^2) of each building.

We start by examining the distribution of activity weights based on the type of aggregation procedure. We observe that the *distance-based* aggregation results in 3525 links between buildings and street network while the *complete accessibility* and *normalized accessibility* aggregation share the connections to the street network and differ only in how the weights are assigned to the connected streets. For computational reasons, we consider the maximum of four connections for each building. For the selected study area, we found 3525 primary, 843 secondary connections, 461 tertiary, and 149 quaternary connections (Figure 104). The primary connections are the same as in the case of distance-based aggregation, whereby the secondary, tertiary, and quaternary connections are the links to all directly accessible streets other than the primary within a given distance threshold (in our case, 20m)

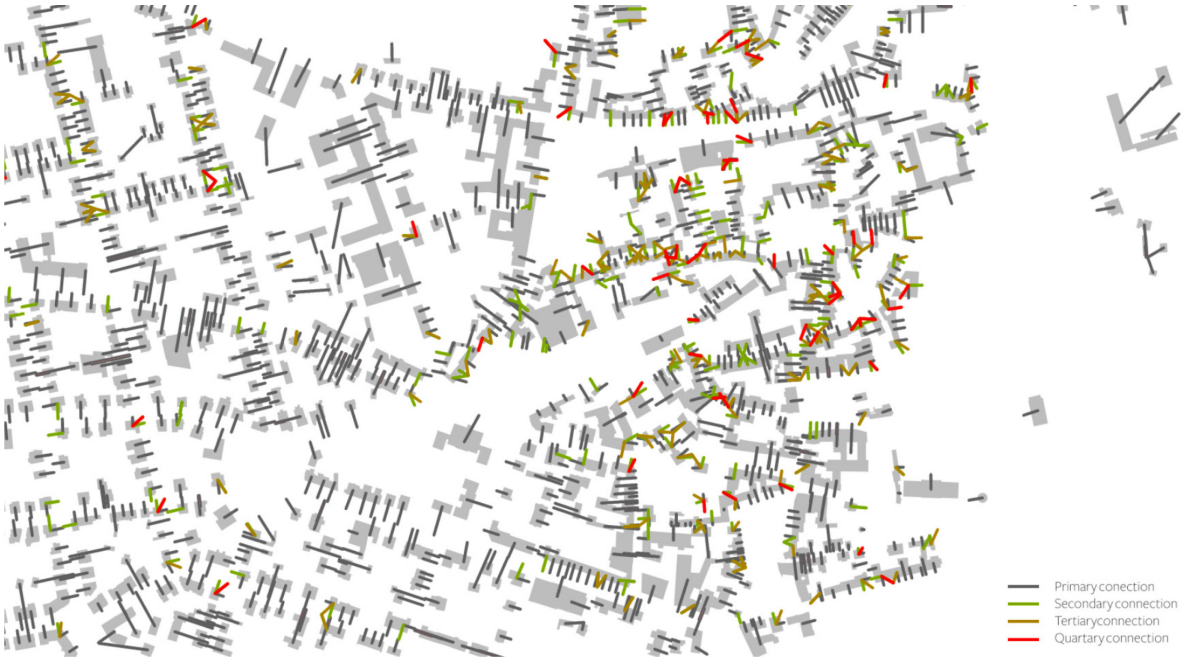


Figure 104. Building to street connections.

Comparing Aggregated Activity Patterns

The total activity weights (total disaggregated weight) is 1.728 e6. After the distance-based and normalized accessibility aggregation cases, the total weight is also 1.728 e6, while the total weight after complete accessibility aggregation is 2.400 e6 (+38% extra weight).

We examine the differences between the individual approaches by correlating the patterns of aggregated weights distributions. We plot the spatial distribution of the difference between street weight calculated by different methods and the respective scatter plots (Figure 105). As expected, we observe that the *complete accessibility* approach produces either equal or higher weightings, but never smaller than the other two approaches (i.e., points in scatter plot are all on or above the 45-degree line). When comparing the *normalized accessibility* and *distance-based* aggregation, we observe a strong linear relationship with no clear spatial pattern of error (i.e., positive and negative residuals are randomly allocated on the map (Figure 105c).

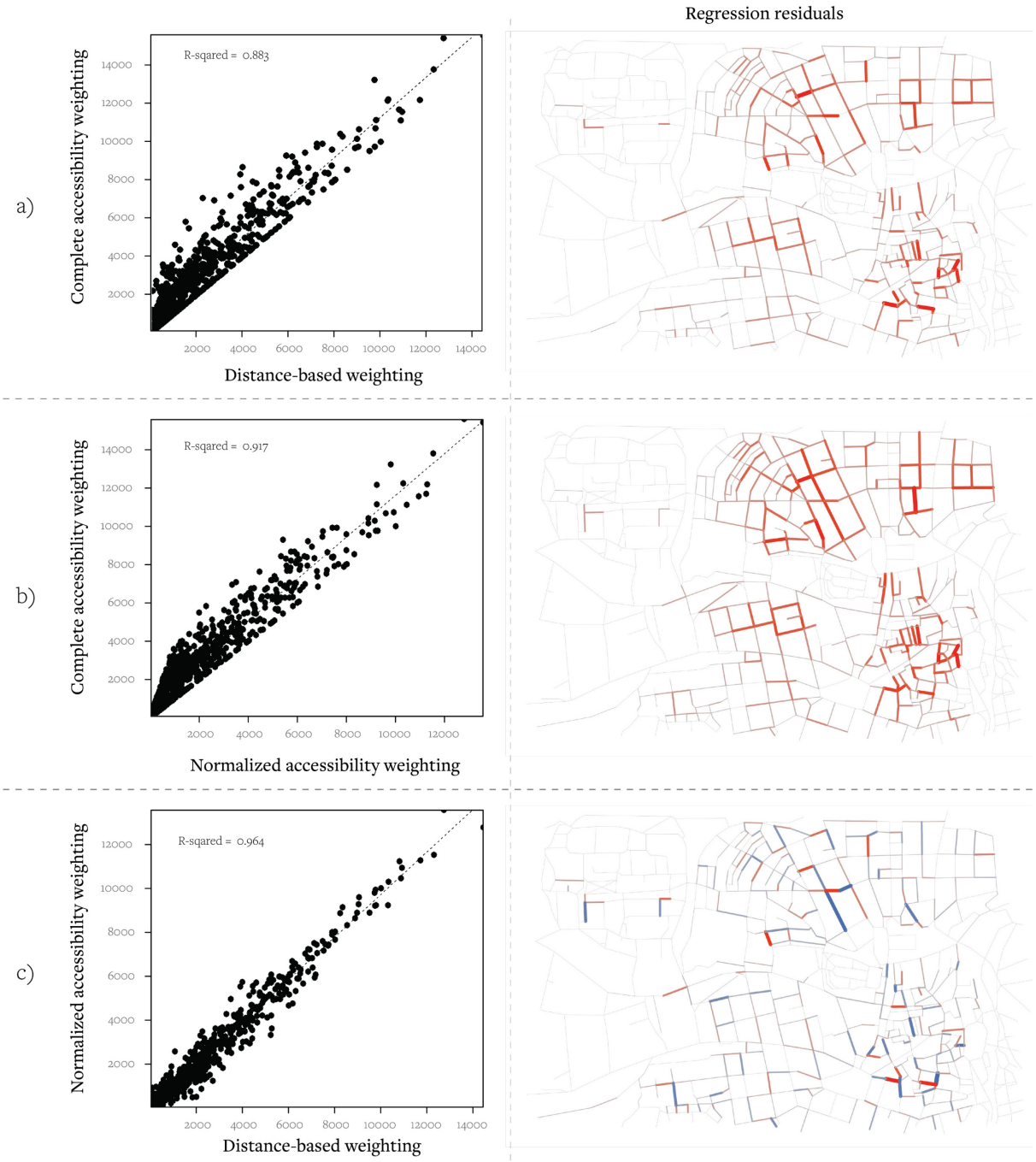


Figure 105. Showing the difference in the aggregated activity patterns produced by the three aggregation procedures. The regression residuals are mapped on a relative scale – the colors and line widths are not comparable between residual maps. Red stands for negative and blue for positive residuals. a) Complete accessibility vs Distance weighting, b) Normalized accessibility vs complete accessibility, c) Normalized accessibility vs distance weighting.

Measuring Movement Model Bias

We calculate the *Through-Frequency* movement[40] based on the three aggregated activity patterns and compare them to the disaggregated model. The origin and destination weights are equal and correspond to the aggregated activity weights calculated by the a) *distance-based*, b) *complete accessibility,* and c) *normalized accessibility* aggregation method.

When comparing the movement flow distribution resulting from the three activity aggregation methods, we conclude that the *distance-based* and *normalized accessibility* method produces almost identical results (R^2 = .99). The biggest difference could be observed between the *complete accessibility* and the two remaining methods. The results are still closely related (R^2 = .97); however, in these cases, the differences (i.e., regression residuals) are spatially structured. We observe the spatial pattern of positive and negative residuals. In the cumulative-based model is the movement frequency in the city center higher in relation to the rest of the city as in the distance-based and normalize-accessibility based model (Figure 106b, c).

[40] The movement model is formally defined as gravitational betweenness centrality with beta = 0.00138 and angular shortest paths..

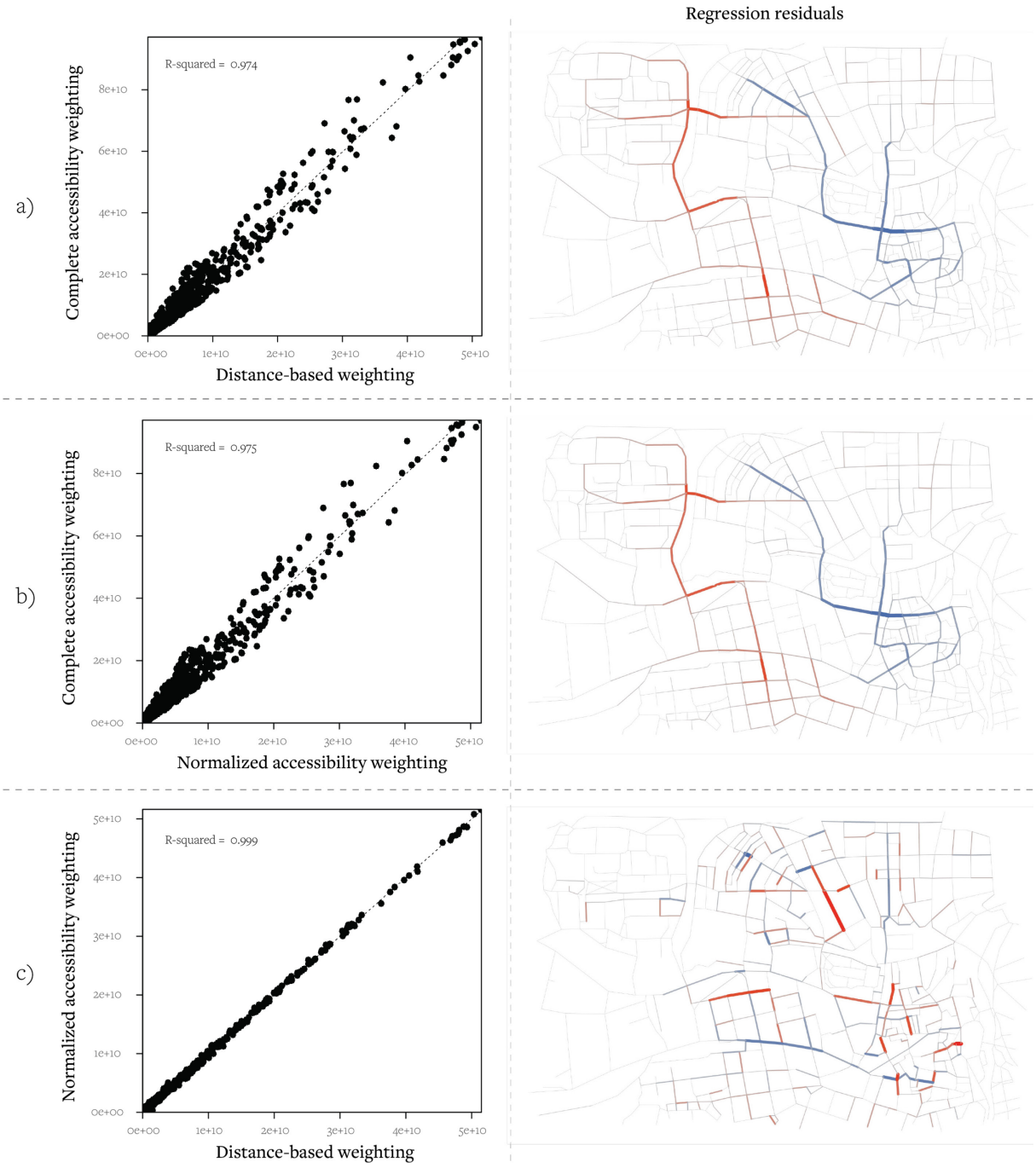


Figure 106. Showing the difference in the movement flow patterns produced by the three aggregation procedures. The regression residuals are mapped on a relative scale – the colors and line widths are not comparable between residual maps. Red stands for negative and blue for positive residuals. a) Complete accessibility vs Distance weighting, b) Complete accessibility vs Normalized accessibility, c) Normalized accessibility vs Distance weighting.

Finally, we measure the bias introduced by each aggregation method by comparing its variance to the disaggregated (i.e., True) movement. For this purpose, we regress the disaggregated movement pattern on each of the movement patterns produced by the three different aggregation methods (Figure 107). The bias is expressed in terms of R^2 - the goodness of fit of the linear model (1 means no bias, 0 means maximum bias). Additionally, we calculate the confidence intervals of the bias via non-parametric bootstrap with N=1000.

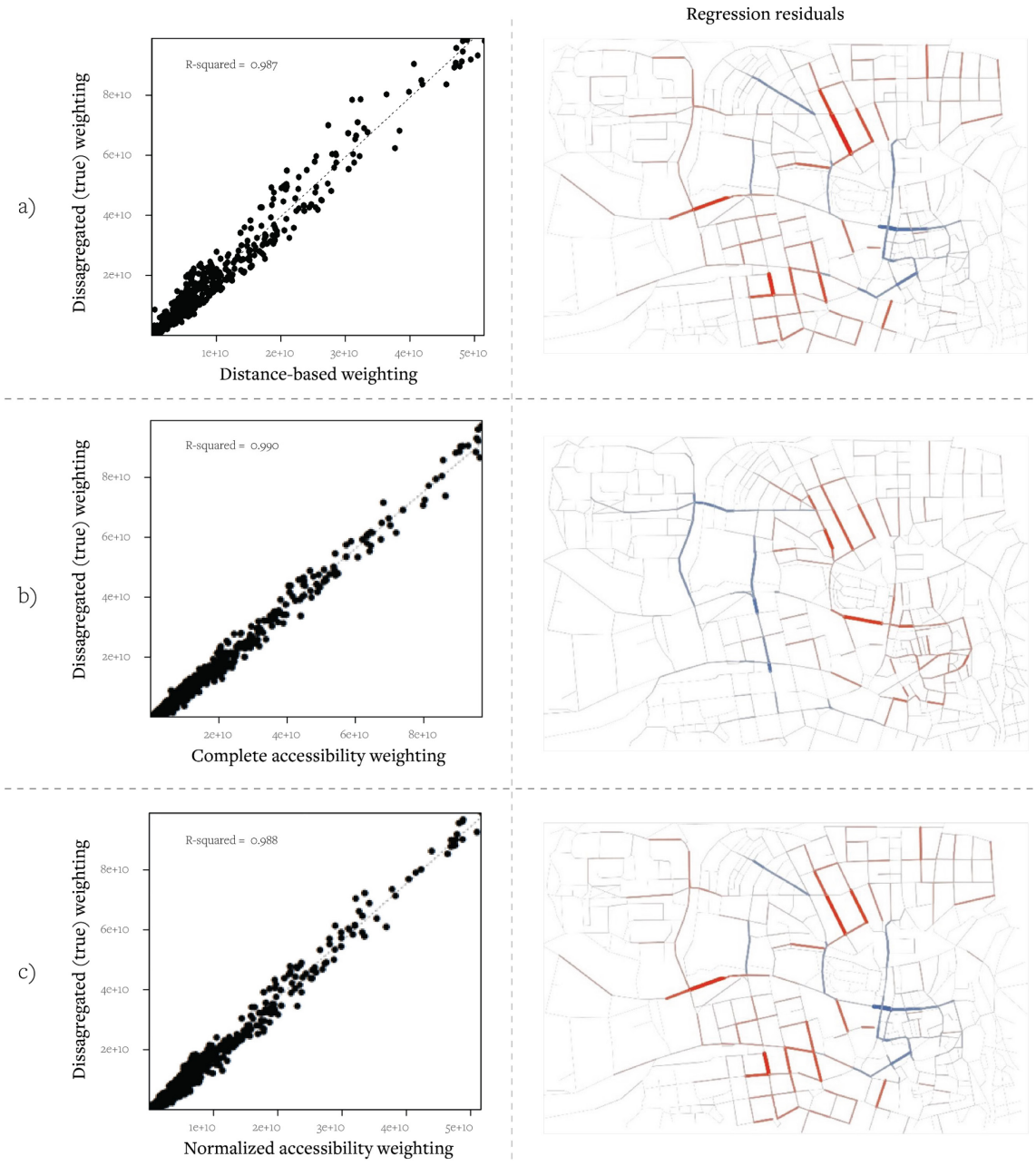


Figure 107. Showing the bias in the movement flow patterns produced by the three aggregation procedures. The regression residuals are mapped on a relative scale – the colors and line widths are not comparable between residual maps. Red stands for negative and blue for positive residuals. a) Distance aggregation vs. true

movement, b) Complete accessibility aggregation vs. true movement, c) Normalized accessibility vs. true movement.

Summary

We found that the bias of *distance-based* aggregation (R^2 = 0.987, CI =0.984, 0.990) is not significantly different from the bias of *normalized accessibility* aggregation (R^2 = 0.988, CI = 0.985, 0.990) and the *complete accessibility* aggregation (R2 = 0.990, CI = 0.987, 0.992). Even though the differences between the methods are negligible when it comes to the pattern as a whole (i.e., what we measure via the R^2), we observe that in the case of few individual streets, the choice of aggregation can make a significant difference. As a result, all three methods can be considered as equivalent when the goal is to model the movement pattern without a special focus on individual locations.

Since, in the case of this study, we focus on the overall pattern, the choice of the method does not significantly influence the accuracy of the movement model and should be driven by other parameters such as computational complexity and interpretability. For this reason, we prefer to adopt the conceptually and computationally simplest out of the three methods – the *distance-based* aggregation.

Appendix 10 Movement Characteristics

From Trip frequency to Trip Volume to Traffic Frequency to Traffic Volume

Movement is complex behavioral phenomenon that involve decisions on where to go and how to get there. In Chapter 2.3, we discuss in detail how to model and simulate movement. Here we address the question of how to represent it. We argue that movement is a multidimensional concept that can be understood only by looking at multiple characteristics at the same time. In the following, we introduce four distinct movement characteristics and empirically test how their pattern and amplitude vary across the study area.

In general, representing movement implies deriving some quantity such as traffic flow, which can, in turn, be compared and put in relation with other quantities at the same location, such as urban form or the allocation of activities. Since quantifying is just another term for measuring, to quantifying movement, we need to address the question of a) where and b) what to measure. Regarding the former, the spatial unit at which all properties (e.g., urban form, activity distribution) are represented throughout this study is the street segment. This being said, we treat each street as one observation for which movement is captured with the consequence that only variation in movement between and not within the street segment can be studied.

Coming to the question of what is measured, we consider who, how often, and how far is moving. These movement characteristics presented here are by no means the complete list and were chosen mainly to answer the research questions stated in this study and to demonstrate the idea of movement as a multidimensional concept.

From vs. Through Movement

We start by considering the individual person who is moving. We characterize the movement at its origin to tell how different locations affect the walking behavior of those who live there. We call it the *From-Movement* as it is capturing the trip characteristics at its origin and considers only those pedestrians who start their trip here. To investigate the effect of walking on the allocation of activities, we assume that these are influenced by all pedestrian traffic passing through a given street regardless of where their trip starts or ends. Accordingly, we call this the *Through-Movement*. In essence, the *Through-Movement* considers all pedestrians, and it is a measure of overall traffic, while the *From-Movement* considers only those who live at a given street segment and captures their individual trip characteristics.

Movement Frequency vs. Volume

As next, we ask two different questions for each pedestrian moving either from or through any given segment. We are either interested in how often they travel or how far they go. The first gives us an idea about the *frequency* and the second about the *volume* of the movement.

Four Movement Characteristics

When combining both aspects together, we end up having four distinct characteristics of pedestrian movement – *From-Frequency*, *From-Volume*, *Through-Frequency,* and *Through-Volume*, which can be classified in two by two matrix (Figure 108). It is important to note that all four movement characteristics should be understood as different but complementary perspectives on the same phenomenon. This becomes clearer when looking at the mathematical model behind each calculation. We see that all four movement characteristics are based on the same trip generation method calculating the likelihood of walking from the origin i to the destination j as a product of the travel demand at origin and travel supply at the destination. In this study, we approximate the distribution of the demand and supply by floor area at the origin and destination of travel. In the case of travel supply or attractiveness of the destination, we also consider how far it is. The applied negative exponential distance decay function is representing how the willingness to travel decreases with the increasing distance.

A closer look at the formal specification of each measure reveals that the *Through-Frequency* is what can be described as a special case of betweenness centrality [41] , and the *From-Frequency* is what is known under the term gravity centrality. The first can be interpreted as a measure of flow and tells us how many pedestrians pass through a given location per unit of time. In essence, the second is the same, with the only difference being that not every pedestrian, but only those who start their trip at a given location are counted. To calculate the *From-Volume,* we multiply the number of trips per unit of time by their total distance. By doing so, we capture the total distance travel by all pedestrians who live at a given location. Finally, the *Through-Volume* is achieved by multiplying the frequency of pedestrians passing through a given street segment by its length.

[41] In this study we use instead of the traditional Boolean distance function (i.e. hard threshold defining what is considered) the negative exponential function (i.e. continuous distance decay penalty) as commonly used in the gravity centrality measure. The resulting centrality measure is therefore variation on the traditional betweenness centrality. We termed it here gravitational betweenness.

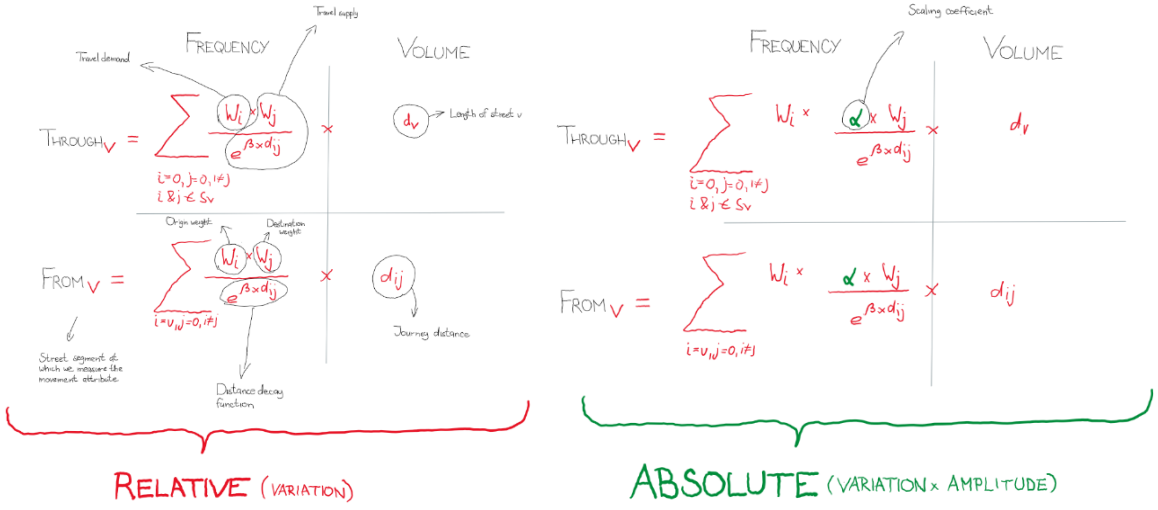


Figure 108. Relationship between different movement characteristics and how this can be used in the process of transforming the simulated movement from relative to an absolute scale.

By now, we discussed the differences between the four movement characteristics; however, we consider it equally important to understand how they relate to each other. Based on their formal definition, we recognized two structural relationships.

First, the *From-Volume* and the *Through-Volume* might be different at the level of individual street segments but are equal in the sum. In other words, the movement volume for the whole system represents the total distance travel in the whole system and can be calculated either by adding up the distances (i.e., volumes) traveled from each segment or distances traveled through each segment.

As next, we discuss the relationship between *From-Frequency* and *From-Volume* (Figure 109). We begin with an individual exploration of each movement characteristic and its dependence on the distance to the destination of movement. In the case of the *From-Frequency*, we observe how the number of trips drops by increasing distance to destination (Figure 109a). In other words, as the distance of the destination grows, our willingness to travel decreases toward zero. The speed of this decrease is controlled by the distance decay coefficient beta. In the case of the *From-Volume*, we observe a more complicated relationship with the distance of the destination (Figure 109b). At nearby destinations, the total travel distance (i.e., *From-Volume*) is also low despite the high number of trips (i.e., *From-Frequency*). As the destination gets further away, the total travel distance increases until reaching its peak. After this point, the trend reverse, and increasing the destination distance reduces the total *From-Volume*. We interpret this nonlinear behavior as a result of two forces pointing in the opposite direction. As we increase the distance of the destination, a) the willingness to travel drops (i.e., *From-Frequency*), which in turn reduces the *From-Volume*, b) however, as we need to travel further to reach the destination, we increase the *From-Volume*.

When we bring the *From-Frequency* and *From-Volume* on the same plot (see Figure 109), we can observe the direct interaction of the above-mentioned forces. It becomes clear that as the frequency drops, the volume is rising until reaching equilibrium of both forces. From here, it starts dropping as the destination distance increase. As a consequence, two different *From-Frequencies* (i.e., different number of trips) can result into the same *From-Volume* (i.e., total travel distance). Thus, these two are structurally different characteristics and must be considered simultaneously. In other words, just by looking at frequency, we know very little about the volume.

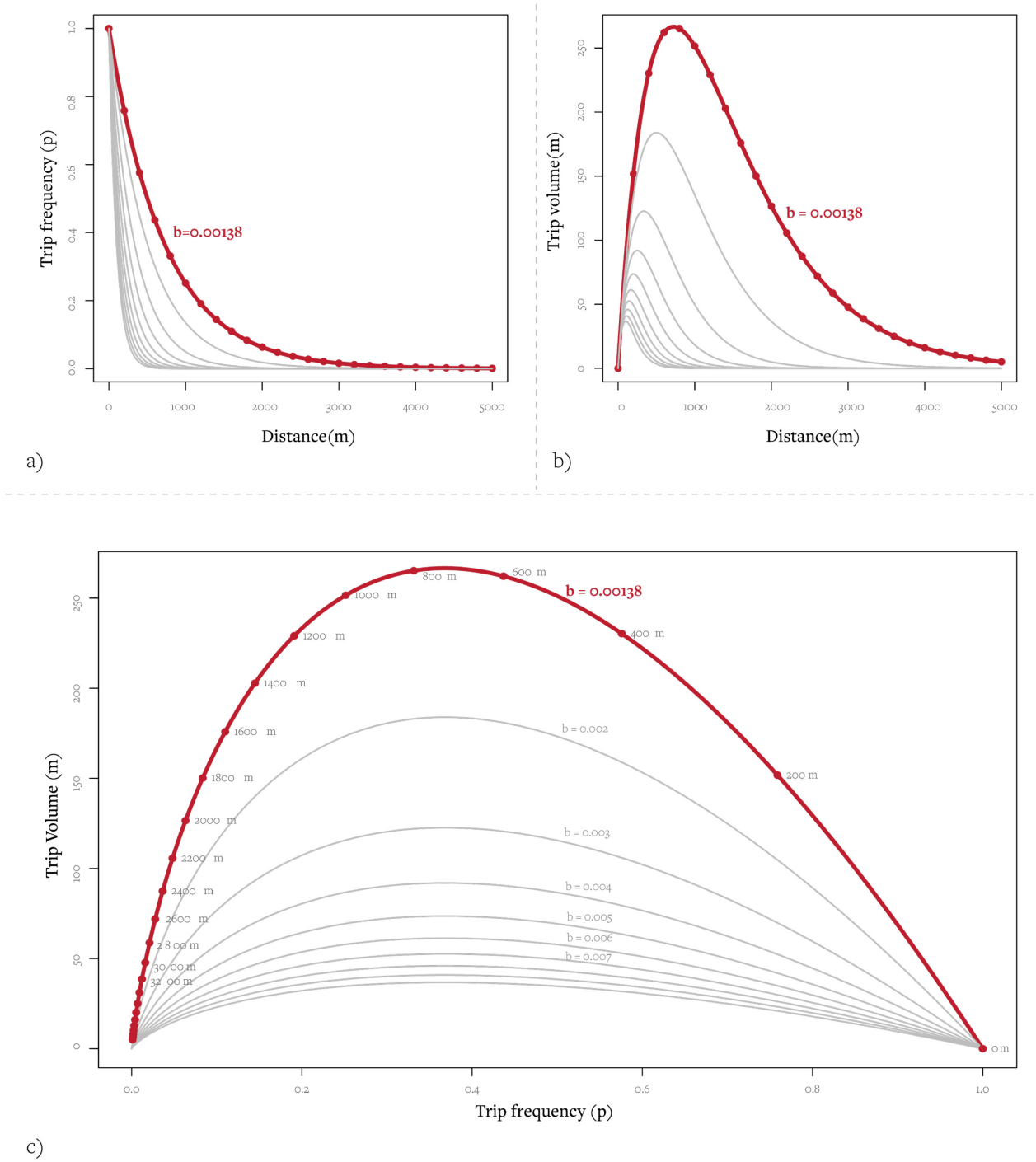


Figure 109. Structural relationship between the movement volume and movement frequency as function of distance to the destination. a) Relationship between trip frequency and the distance of the destination. b) Relationship between trip volume and the distance of the destination. c) Relationship between trip frequency and trip volume. Black curve is calculated for the distance decay coefficient $\beta = 0.00138$, grey curves are representing the same function at $\beta = 0.002$; .003; .004; .005; .006; .007; .008; .009; .01

Appendix 11 Relative to Absolute Movement

In the following, we discuss in detail how to estimate the missing information on amplitude and variance in the movement components pyramid (Figure 110). The point of departure is the empirical data on total movement (i.e., top of the pyramid) and the simulated variance of the movement components for individual activities (i.e., bottom of the pyramid) with the rest of the pyramid remaining unknown.

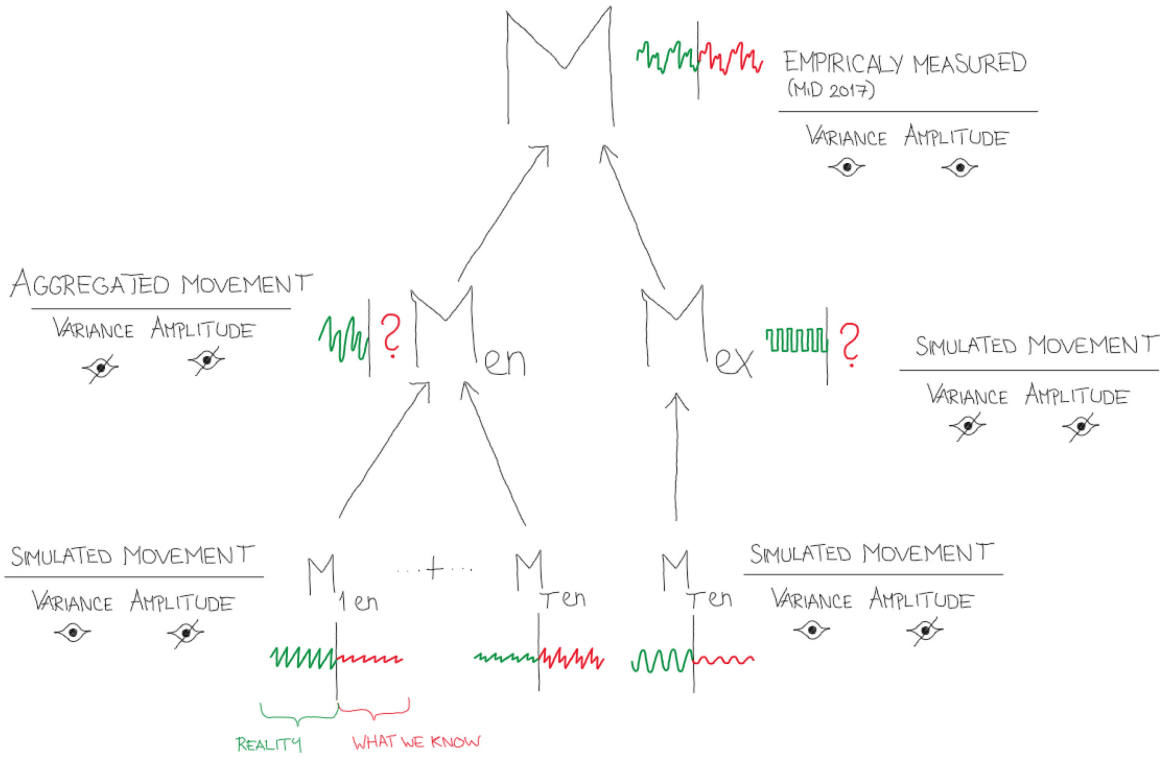


Figure 110. Overview of the knowns and unknowns of the movement components pyramid. The movement pattern is fully known (i.e., can be described by its variance and amplitude) only for the total movement M.

Estimating Variance of Aggregated Exogenous and Endogenous Movement

In the case of exogenous movement components for individual activities $M_{T(ex)}$, the problem of estimating the combined M_{ex} is straight forward. Since they all are based on the same random process in activity allocation, the resulting movement is also following the same variation pattern. Coming back to our audio signal metaphor from Chapter 4.2.3.1, it is the same as having multiple versions of the same audio track in different levels of loudness. In such a case, if we combine them, we still end up having the same track in terms of its wavelength regardless of the loudness of the individual components. In other words, the singer will still be singing the same song. To formalize this, we express the relationship

between the $M_{T(ex)}$ and M_{ex} in mathematical terms. We split the pattern in the known variance represented by the vector M and the unknown amplitude represented by the constant α.

$$M_{ex} = \alpha_1 M_{1(ex)} + \alpha_2 M_{2(ex)} + \cdots + \alpha_T M_{T(ex)} \quad (34)$$

The key point here is that the individual exogenous movement component for each activity type $\alpha_T M_{T(ex)}$ are linear multiplications of the same vector.

$$\alpha_{(1,T-1)} M_{(1,T-1)(ex)} = \beta_{(1,T)} M_{(T)(ex)} \quad (35)$$

As a result, we substitute and simplify the relationship described in Equation 34 through a single term.

$$M_{ex} = \beta_1 M_{T(ex)} + \beta_2 M_{T(ex)} + \cdots + \alpha_T M_{T(ex)} \quad (36)$$

$$M_{ex} = \beta_{ex} M_{T(ex)}\,;\ \beta_{ex} = \beta_1 + \beta_2 + \cdots + \alpha_T \quad (37)$$

From this, we directly conclude that the variance in M_{ex} is equal to the variance in any of the $M_{T(ex)}$ components with the only unknown being the amplitude.

Estimating Amplitude of Aggregated Exogenous and Endogenous Movement

In the case of the endogenous movement components for individual activities $M_{T(en)}$, these cannot be simply combined to M_{en} as they differ in terms of their variation. Coming back to our metaphor, these are different audio tracks and cannot be combined without proper scaling of their loudness. Otherwise, the result would not be the same song anymore. In mathematical terms, we can express this relationship similarly as in the exogenous case:

$$M_{en} = \alpha_1 M_{1(en)} + \alpha_2 M_{2(en)} + \cdots + \alpha_T M_{T(en)} \quad (38)$$

with the difference that the individual components are not the multiplication of the same vector.

$$\alpha_{(1,T-1)} M_{(1,T-1)(en)} \neq \beta_{(1,T)} M_{T(en)} \quad (39)$$

To scale the endogenous movement components for different activities in proper relation to each other, we use the MiD2017 travel data capturing the number of trips attracted by each travel activity (Appendix 7). Unfortunately, the MiD2017 data is collected for the whole

study area and not for the individual street segments. This means that we can use the proportions found for the whole city of Weimar to scale the $M_{T(en)}$ components in proper mutual relationships; nevertheless, the absolute value of the estimated movement will still remain unknown (Figure 111). After the scaling procedure, the relationship between the individual endogenous movement components and the aggregated endogenous movement can be described as following:

$$M_{en} = \beta_{en}(M_{1(en)} + M_{2(en)} + \cdots + M_{T(en)}) \tag{40}$$

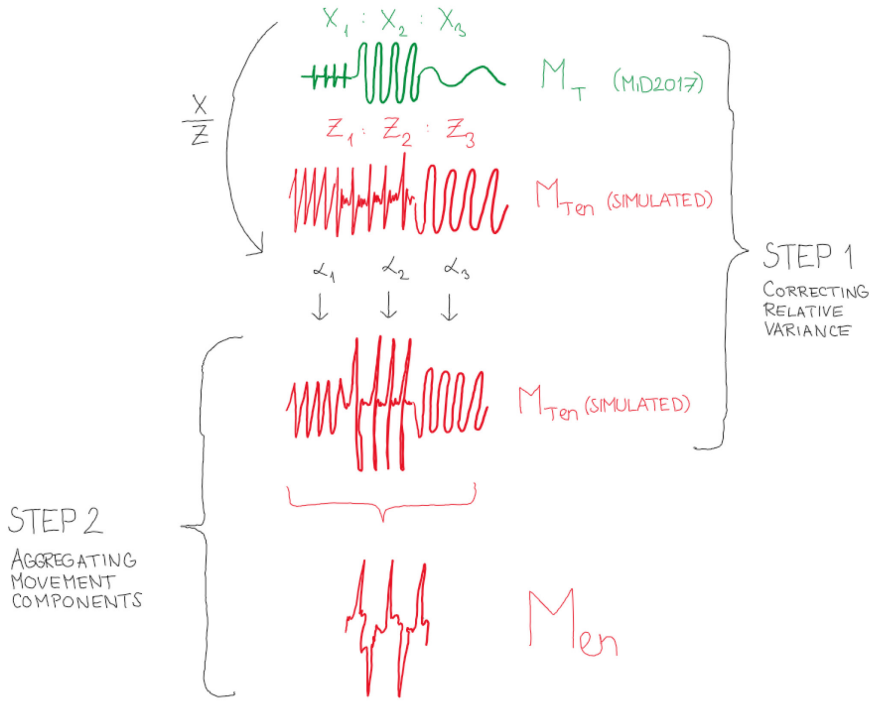


Figure 111. Correcting the relative proportions of the simulated endogenous movement pattern by activity $M_{T(en)}$ (step 1) and combination of the corrected movement components to the aggregated endogenous movement M_{en}.

To summarize, at this point, we were able the deduce the variance in all movement components while the amplitude remains unknown for all but the total movement (Figure 112).

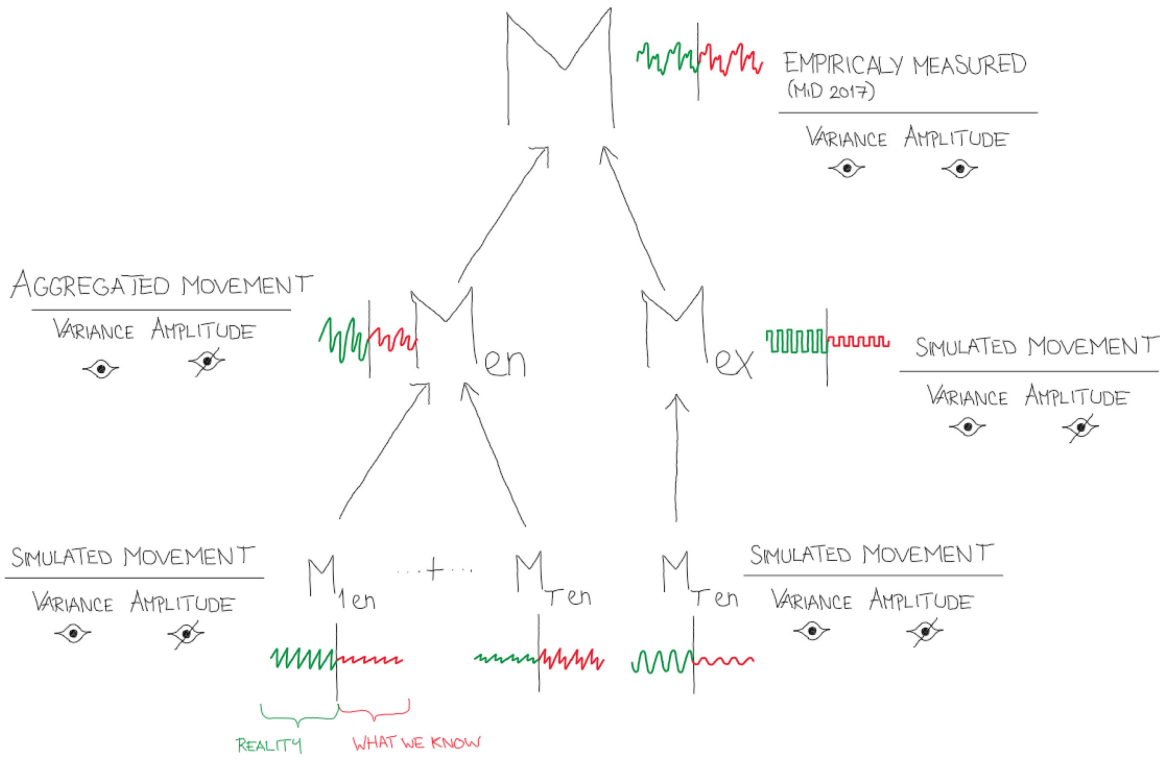


Figure 112. Overview of the knowns and unknowns of the movement components pyramid.

Estimating Amplitude of Exogenous and Endogenous Movement by Activity Type

To estimate the contribution of endogenous and exogenous movement component to total movement, we estimate the portion of the variance in the total movement explained by each movement component. Explaining variance in one continuous dependent variable by multiple explanatory variables is a standard procedure accomplished by the linear regression model (see Equation 41).

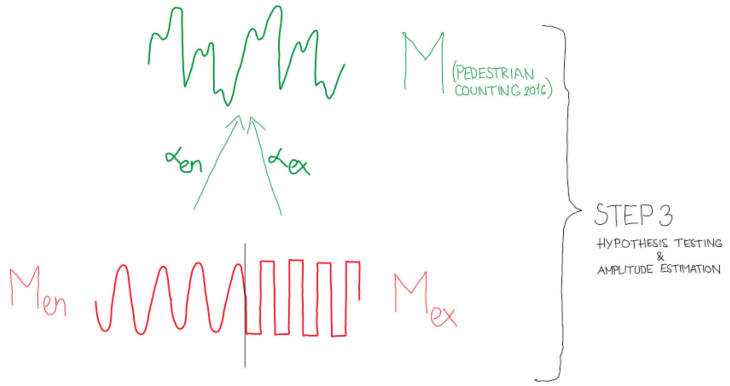


Figure 113. Estimating the absolute amplitude of the aggregated endogenous and exogenous movement pattern M_{en} and M_{ex}.

$$General\ case: y = \alpha_1 x_1 + \alpha_n x_n + \varepsilon; \quad (41)$$

$$m_{total} = \alpha_{ex} m_{exogenous} + \alpha_{en} m_{endogenous} + \varepsilon \quad (42)$$

The $\alpha_{ex}$and α_{en} coefficients are formally interpreted as the expected increase in total movement associated with one unit increase in the simulated movement components. In essence, this is telling us how to scale each movement component to explain as much variance in the total movement as possible. To guarantee that estimated coefficients are only positive real numbers as movement cannot be negative, we estimate a special version of linear regression – the penalized regression. The penalized regression makes it possible to restrict the range of each coefficient and thus guarantee that the results are interpretable and match the theoretical constraints (James et al., 2014; Bruce and Bruce, 2017). It must be noted that penalty constraints apply only to the estimated coefficients and not to the error term. The error term represents the movement that was not captured by our simulation model (e.g., additional activities). The consequence of the penalty regression specification is that this error term is assumed to take both positive and negative values, which still present conceptual difficulty. Ideally, we would need a model assuming that both the coefficients and error are only positive. Nevertheless, to our best knowledge, such an estimator is currently not known, and therefore we consider the penalized regression as the best available alternative.

Finally, we substitute the estimated scaling coefficient for M_{en} in the Equation 38 to find the amplitude of the $M_{T(en)}$ components (Figure 114).

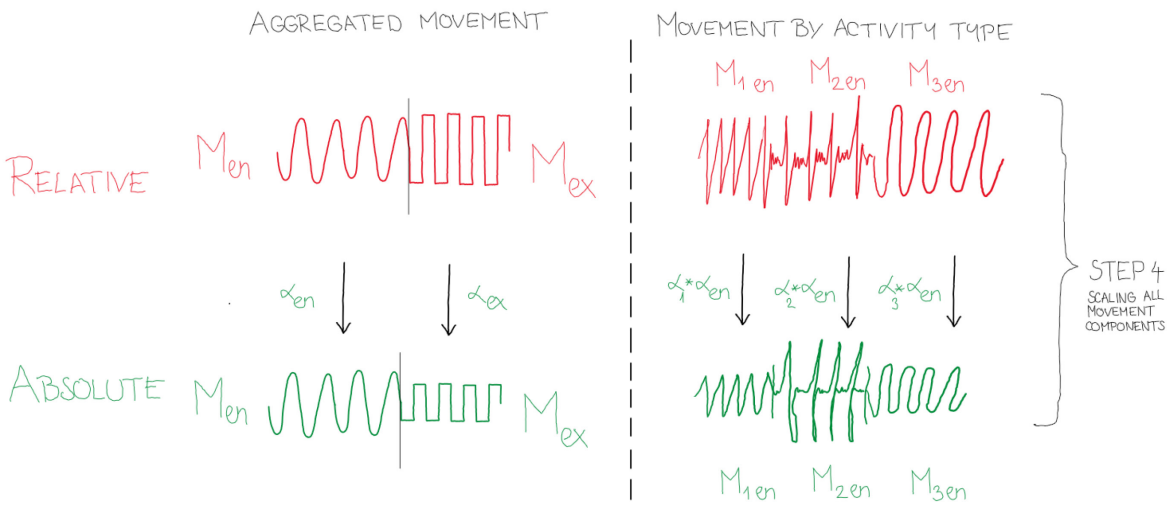


Figure 114. Scaling the amplitude of all movement components.

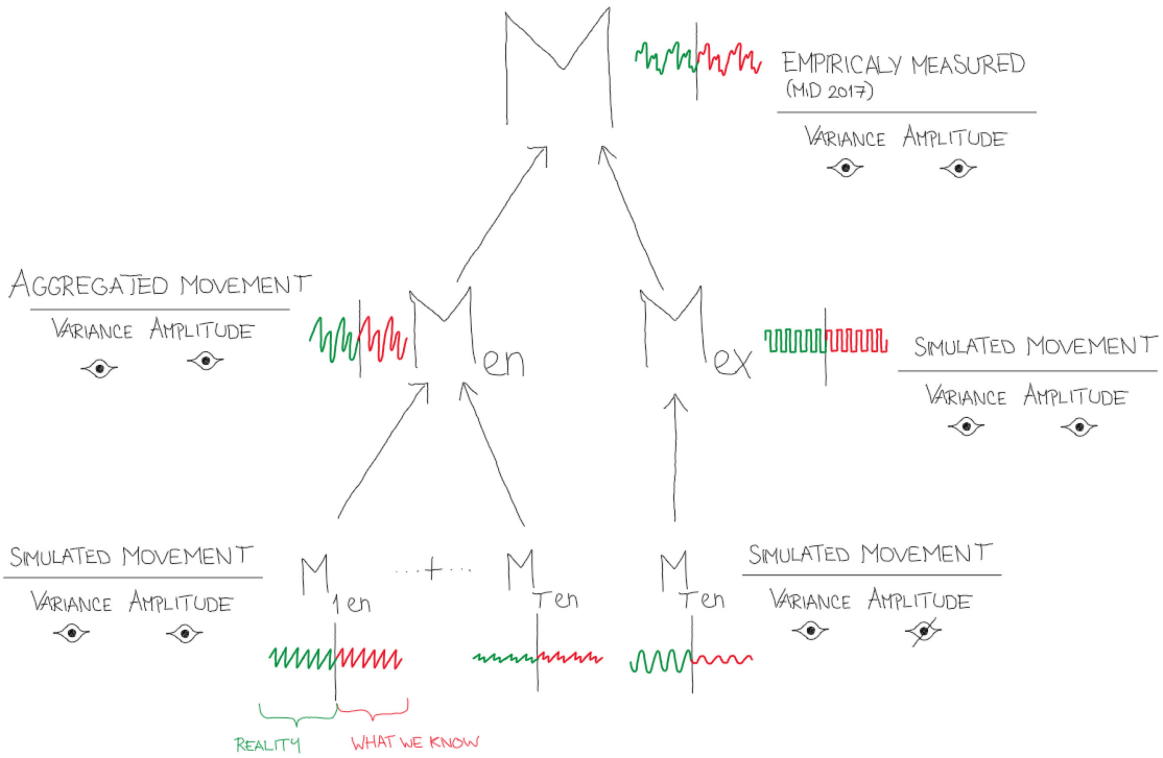


Figure 115. Movement pyramid after the estimation process. The variance and amplitude of all variables, with the exception of the exogenous movement by activity, is fully known.

Appendix 12 Separating Structure from Noise

As discussed in Chapter 2.5.1, the variation in the simulated $M_{T(en)\,sim}$ is the combination of $M_{T(en)}$ - the movement component derived from the allocation of activities and $M_{T(ex)}$ - the random movement component directly derived from the urban form. Formally, we express this relationship as $M_{T(en)\,sim}$ being a linear combination of the unknown $M_{T(en)}$, the known $M_{T(ex)}$ (from simulation) and the error term μ.

$$M_{T(en)\,\mathrm{sim}} = \alpha_{en} M_{T(en)} + \alpha_{ex} M_{T(ex)} + \mu \tag{43}$$

$$M_{T(ex)\,\mathrm{sim}} = M_{T(ex)} \tag{44}$$

$$M_{T(en)\,\mathrm{sim}} = \alpha_{en} M_{T(en)} + \alpha_{ex} M_{T(ex)\,sim} + \mu \tag{45}$$

The challenge here is to identify the endogenous movement component $M_{T(en)}$ when the α_{en}, α_{ex} and μ are unknown and only the information on $M_{T(en)\,sim}$ and $M_{T(ex)\,sim}$ is available. In other words, we try to separate the structural (i.e., endogenous) and random noise (i.e., exogenous) movement components from their joined distribution. We must note that since both, the α_{en} and $M_{T(en)}$ are unknown, we can only identify their product. Therefore, the actual challenge is to identify $\alpha_{en} M_{T(en)}$.

We can imagine the task as having a song composed of two tracks A and B. We have the information about the song and track A, both at an arbitrary level of volume (Figure 116). The task is to get the track B. The error term μ represents here some random noise in the song, which cannot be attributed to any of the two tracks. The task is to find the right volume level of the known track. A so we can subtract it from the song to get the track B.

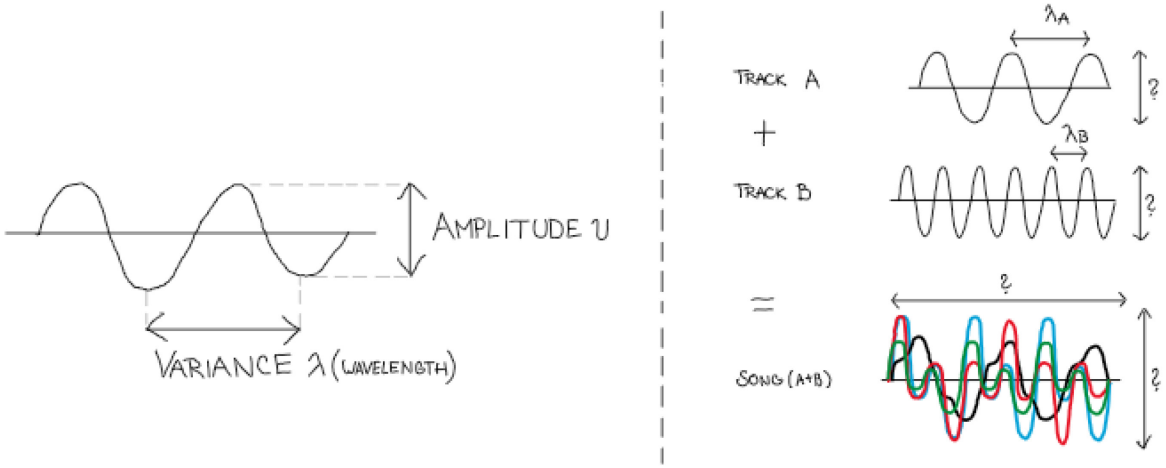


Figure 116. Combining two patterns with unknown amplitude illustrated on the example of sound waves. Without the information on the amplitude, results with different variances are possible.

We apply the same principle to the original problem and systematically search for the proper α_{ex} coefficient to subtract $\alpha_{ex}M_{T(ex)\,sim}$ from $M_{T(en)\,sim}$ to get $\alpha_{en}M_{T(en)}$. We let the α_{ex} grow and select the optimal value extracting maximum variance from the $M_{T(en)\,sim}$ by producing minimal error. The error is introduced when α_{ex} the coefficient is too large, and the outcome of the resulting $\alpha_{en}M_{T(en)}$ movement is negative (Figure 117), while the extracted variance gets larges as the α_{ex} coefficient grows. For this reason, we are looking for a trade-off between the variance maximizing and error minimizing function of α_{ex}.

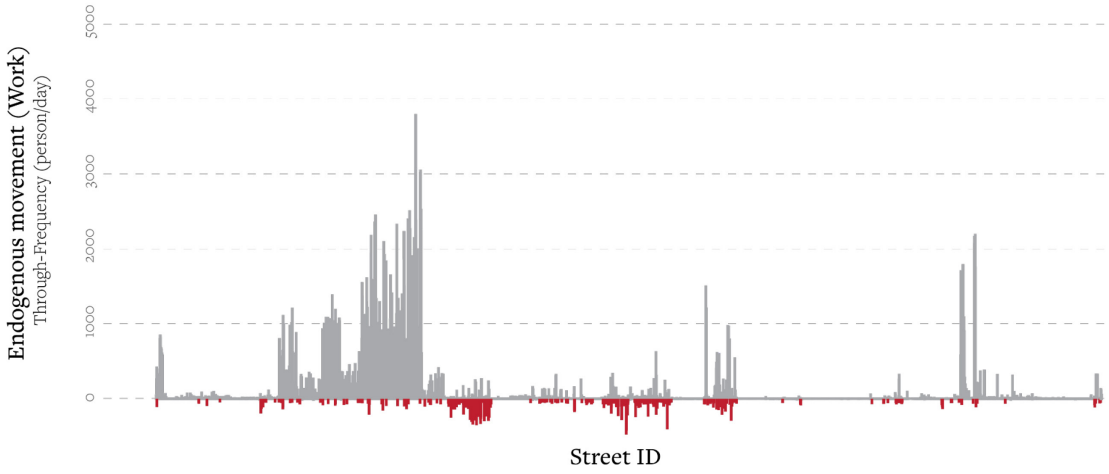
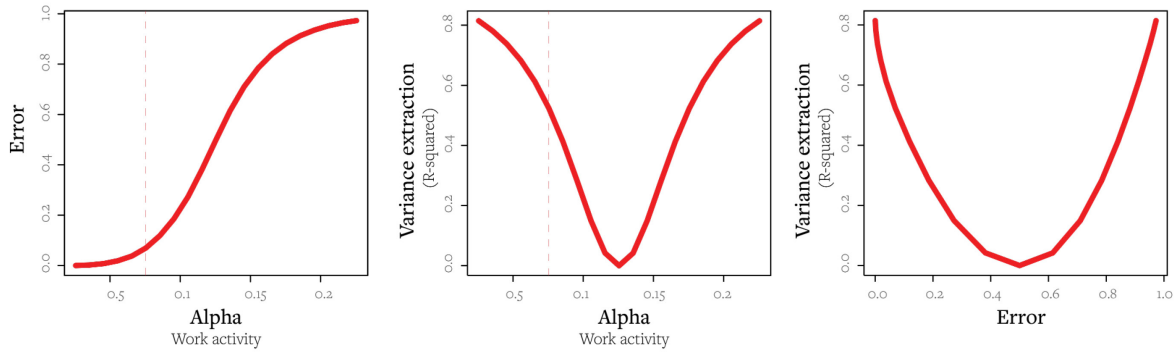


Figure 117. Example of error - negative movement in the extracted endogenous movement. We visualize the variance in the movement as a bar plot with one bar for each street segment. The negative movement (i.e., error) is displayed in red color; the positive movement is displayed in grey color.

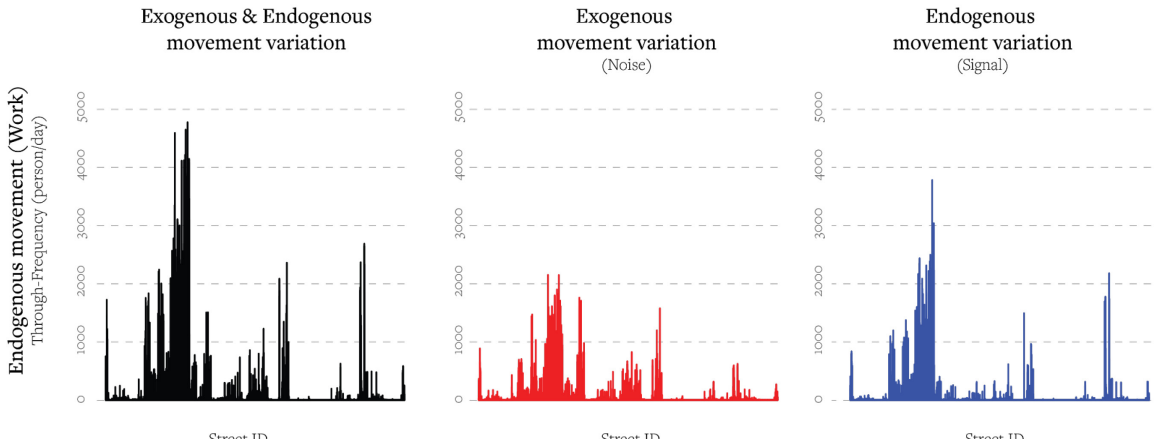
In the following, we exemplary demonstrate the extraction procedure on the work activity type. We measure the relationship between α_{ex} both optimization criteria (i.e., variance and error) from the simulated data on $M_{T(en)\,sim}$ and $M_{T(ex)\,sim}$. As depicted in Figure 118, we observe a) sigmoid error growth curve (Figure 118a) for the error and b) symmetrical inverse bell shape for variance extraction. Both functions have their optimum at value zero, and as can be seen in Figure 118c their mutual relationship is Pareto optimal (i.e., none of the two criteria can be improved without making the other criterion worse). From the Pareto front depicted in Figure 118c, we chose the closest point to the optimum at [0,0] coordinates. This value represents the optimal α_{ex} coefficient to extract the unknown $\alpha_{en}M_{Work(en)}$ (Figure 118).



a) b) c)

Figure 118. Plots of the two optimization criteria for work activity type and α_{ex} coefficient (i.e., scaling factor). a) Function of α_{ex} and error – best is 0 worst is 1. b) Function of α_{ex} and variance extraction – best is 0 (no shared variance has remained) worst is 1(the variance of the $M_{T(ex)\,\mathrm{sim}}$and $M_{T(ex)}$is identical). c) Pareto front – the relationship between variance and error optimization criterion.

We note that the value of the optimal $\alpha_{Work\,(ex)} = 0.75$ does not have any substantial meaning as it depends on the arbitrary chosen constant exogenous activity allocation pattern $A_{Work(ex)}$ driving the simulated exogenous movement $M_{Work(ex)\,sim}$.



a) b) c)

Figure 119. Visualization of the variance in the movement as a bar plot with one bar for each street segment. a) Is the known $M_{Work(en)\mathrm{sim}}$ (Structure & Noise). b) Is the known $M_{Work(ex)\,\mathrm{sim}}$ (Noise) multiplied by the fitted $\alpha_{Work(ex)} = 0.75$. c) Is the extracted $\alpha_{en}M_{Work\,(en)}$ (Structure) with all negative values (extraction error) equal to zero.

Appendix 13 Defining Distance

Traditionally, movement modeling is based on the aggregation of paths between pairs of nodes (origins and destinations) in the street network. These paths are assumed to be “the result of minimizing procedures such as selecting the shortest path, the quickest path or the least costly path." (Golledge, 1995, p1.). This assumption is based on the least effort principle, which, as Zipf puts it, “is a broad theory that covers diverse fields from evolutionary biology to webpage design. It postulates that animals, people, even well-designed machines will naturally choose the path of least resistance or effort" (Zipf, 1949, p68.). In the literature, the term shortest path is often understood in a more general sense representing routes optimal for any criteria defined by the person undertaking the journey. One can try to minimize the distance or the time spend during the travel and maximize safety, aesthetic qualities of the path or even combine and optimize for multiple criteria at the same time. As a result, based on the optimization criteria, there will be different shortest paths between the same origin-destination pair.

As briefly mentioned in the literature review on pedestrian movement models (Section X), different scholars advocate for different metrics for calculating travel costs. The most prominent travel distance metrics the metric or Euclidian (Porta, 2010; A Sevtsuk, 2010), cognitive - angular (Hillier & Iida, 2005; Turner & Dalton, 2005) or topological (Hillier & Hanson, 1984b) distances. Before going into any details, it must be noted that regardless of the optimization criteria, we always assume the prior knowledge of the environment. This is important as all approaches mentioned here are limited to modeling the navigation in a familiar environment. In the context of our study design, this means that pedestrian behavior of individuals such as tourists who do not navigate but rather explore is not captured by this type of model.

Physical Distance

The first and widely used route optimization criterium is based on the assumption that people tend to minimize the energy necessary to reach the destination. This definition of distance comes from the realm of transportation research and comprises aspects such as Euclidian distance, time, or slope (Garrison, 1960; Kansky, 1963). The Euclidian distance is closely related to travel time even though “time is a better measure of convenience compared to path length, when it comes to measuring distances for pedestrians and cyclists” (Nourian et al., 2015, p10). The crucial factors accounting for the differences between metric and temporal paths are a) topography, b) presence of under and overpasses, and c) street crossing waiting times (Rodríguez & Joo, 2004; Troped et al., 2001). Despite these differences, we consider both, time and distance as measures capturing the physical effort required to reach the destination and thus belonging to the same category. The clear advantage of the physical distance-based approach to path selection is its universal

applicability and ease to operationalize. This is to say that the physical distance will remain the same in Munich (Germany) as well as in Peking (China) and can be easily measured in meters, miles, or other well-established units.

Cognitive Distance

In contrast to the physical distance, the second category of shortest path optimization criteria is based on the aim to minimize the mental effort. Here the assumption is that the path chosen to reach the destination has to be created from the mental representation of the environment – cognitive map (Tversky, 1993) and that people tend to choose paths which easy to construct and navigate (Gale et al., 1990; Golledge, 1995). In general, we can say that "routes involving a greater number of turns are more taxing on memory and therefore harder to integrate into cognitive maps" (Sevtsuk, 2010, p49). Space syntax scholars are the most prominent advocates of the cognitive approach to modeling movement and brought it to the common planning practice. The established operationalizations of this approach are the topological distance (Hillier, 1984) and its later refinement termed as the angular deviation (Turner, 2001).

The general idea for both measures of cognitive distance is grounded on the advances of cognitive psychology and aims to quantify the cognitive complexity of a route. In the case of topological distance, this is simply the number of path segments – topological steps of which a path consists of. The alternative approach is to count the angular deviation between two neighboring street segments. This approach does not rely anymore on the ambiguous axial map while keeping its core idea – routes on straight or close to straight lines are easier to navigate. Since long axial lines can be divided into smaller segments without affecting the analysis, it offers a more refined and robust approach. One of the most profound consequences is that the angular shortest path can be computed on the basis of the established and widely accessible representations of street networks such as a street-center line with only little adjustment necessary (Krenz, 2017).

Mixed and Alternative Models

Even though the arguments for both, the physical and cognitive approach to path selection present a convincing argument, empirical research shows that people are neither choosing their routes purely on the basis of Euclidian distance nor exclusively on the basis of cognitive effort. Given that "both angle and distance seem to correspond to influence traffic flow, the two methods of route choice might even be combined to reflect genuinely cognitively shortest paths for different levels of knowledge of the system" (Turner, 2007, p17). Even though we do not fully understand how is the physical and cognitive distance combined, the empirical evidence suggests that the chosen path is as straight as possible (Dalton, 2001), but also not

more than 20% longer than the Euclidian shortest path (Li & Tsukaguchi, 2005; Takeuchi, 1977).

In spite of the empirical evidence favoring the mixed model, there are several unresolved questions hindering its implementation. It remains unclear how to combine the different criteria (e.g., meters and angles), and this would require new algorithms to search street networks for the optimal path effectively. Thus, we consider the mixed model as a developing field of research and promising future alternative. However, we argue that for the time being, it does not represent a feasible alternative to the established physical or cognitive computational route choice models.

Overall, it is clear that all current computational models discussed here are a crude simplification of human wayfinding. For example, in the case of human cognition, the path selection is guided by complex processing of inaccurate mental representation of the world (Frankenstein, 2015), the computational models discussed here are rather based on simple processing of very accurate representations of reality. This sharp contrast was among others expressed by Golledge, who argues that it is unclear if shortest path calculations "are the criteria used by humans or are they methods best suited to the mathematical determination of optimal paths through complex multi node networks to ensure economic efficiency of commercial traffic fleet, but yet using criteria of which people are in general unaware or are incapable of using?" (Golledge, 1995, p2). In other words, there is little evidence proving that humans are capable of an accurate assessment of Euclidian distances or angular deviation between origin and destination of travel. So even if we try to optimize our paths for such criteria, the results are expected to be biased and sub-optimal. In this context, we argue that a more probabilistic approach allowing for range of alternative paths within a range to the optimal solution might be more realistic model of human wayfinding.

Pedestrian Route Choice – Empirical Study

This being said, we still consider the current physical and cognitive models of path selection as useful, even though they might be improved in the future. Nevertheless, it is difficult to pick one, as neither theory nor the empirical data offer a clear answer to the question of which one of them matches better human wayfinding. Since this might depend on the individual characteristics of the urban fabric, we conduct an empirical study in the city of Weimar comparing the ability of physical and cognitive shortest path to capture pedestrian movement (Appendix 7). Based on the results of the Shortest Path Study 2017 (SPS2017), we select the best performing option and use it throughout this research to estimate pedestrian path selection and travel behavior.

Participants of this study (N=50) were students with a high degree of familiarity with the case study area (longer than one year). All participants were asked to keep a detailed travel

diary recording each journey with map drawing capturing the path taken, travel purpose, time of departure, and transportation mode. Our methodology was based on the MiD2017 study capturing multi-purpose journeys as individual routes with their respective travel purpose. In total, we collected 529 routes for four transportation modes (Walking, Cycling, Bus, Car) and seven[42] travel purpose categories (Accommodation, Work, Education, Shopping, Administrative, Healthcare, Gastronomy) matching the activity types adopted in this research (see section X).

To test which computational model of path selection – the physical or cognitive, captures best the behavior of pedestrians in Weimar, we quantify how closely they match the empirical path. The matching accuracy of the computational path is measured as a percentage of the computational path length from the total path length of the empirical path. The resulting measure of accuracy is ranging from zero (i.e., no match) to one (i.e., perfect match) for each path. Finally, we calculate an average accuracy score per model type and compare the overall scores. Accordingly, we select the best performing model for the computation of pedestrian paths and utilize it throughout this study.

Provided that the scope of this research is limited to walking, we select from the complete data set the pedestrian routes (N = 369) and use these for the shortest path model selection. Before running the model selection analysis, we deal with duplicate paths in the data set. These emerge when participants take the same path multiple times throughout the study period or walk along the same path in the opposite direction. Including the duplicate paths in the data set would result in giving more weight in the model selection procedure to the paths which are taken more often. We consider this approach as valid as long as the travel behavior captured in the study is representative of the general population in Weimar. Or, in other words, if pedestrians in Weimar take some paths more often than others, it is valid to give more weight to these paths.

To assess the representativeness of our sample consisting of 50 students, we compare the travel behavior recorded in this study to a representative random sample of the population travel behavior captured by the MiD2017 study (see Appendix 7). In particular, we compare the distribution of a) travel modes and b) the travel purpose recorded in both studies. For each comparison, we calculate Pearson's Chi-squared test, which evaluates how likely it is that the selection of participants influenced the distribution of travel modes and travel purposes. Suppose the null hypothesis of the Chi-squared test is that the selection of participants does not affect the distribution. In that case, the likelihood below 5% (i.e., p-

[42] We combine the six destination activities ant accommodation as origin activity. Since the route choice algorithm is symmetric (i.e. the same route is chosen from A to B as from B to A) for purpose of the route choice study the origins and destinations are interchangeable.

value < 0.05) means that we must refuse the null hypothesis and consider the results of Path study 2017 as representative.

Table 19. Distribution of transportation modes for the SPS2017 and MiD2017 study

	Walking	Cycling	Public transport	Car
MiD 2017	248 (39%)	89 (14%)	61 (9%)	233 (39%)
PathStudy 2017	369 (70%)	39 (7%)	111 (21%)	5 (1%)

A chi-square test of independence was performed to examine the relationship between study design (MiD2017, PSP2017) and the distribution of travel mode. The relation between these variables was significant, $X^2 = 125.21$, df = 3, p = <.001. As a result, we refuse the null hypothesis in favor of the alternative - the choice of participants affects the distribution of travel modes.

Table 20. Proportion of trips by travel purpose for PSP2017 and MiD2017 study.

	Work	Education	Shopping	Free time	Home
MiD 2017	20 (8%)	6 (2%)	11 (4%)	28 (11%)	53 (21%)
PathStudy 2017	12 (2%)	159 (30%)	58 (11%)	66 (12%)	162 (30%)

As next, we perform a chi-square test of independence to examine the relationship between study design (MiD2017, PSP2017) and the distribution of travel purposes. The relation between these variables was significant, $X^2 = 71.49$, df = 4, p = <.001, meaning that the choice of participants affects the distribution of travel purposes.

Based on the Chi-squared test of independence, we conclude that the choice of travel mode and destination were affected by the selection of participants and thus are not representative. Since the PSP2017 cannot be considered as a representative, we evaluate only unique pedestrian paths. This results in 203 paths after removing 166 duplicates which are used in the model selection procedure (Figure 120).

Figure 120. Set of pedestrian paths as captured by the Path selection 2017 study. Paths are laid on top of each other, with the color intensity indicating multiple paths sharing the same street segment.

As first, we calculate for all 203 journeys the shortest physical and cognitive paths. Due to the reasons discussed above, we adopt the Euclidian distance for the former and the angular deviation as optimization criteria for the latter (see Figure 121).

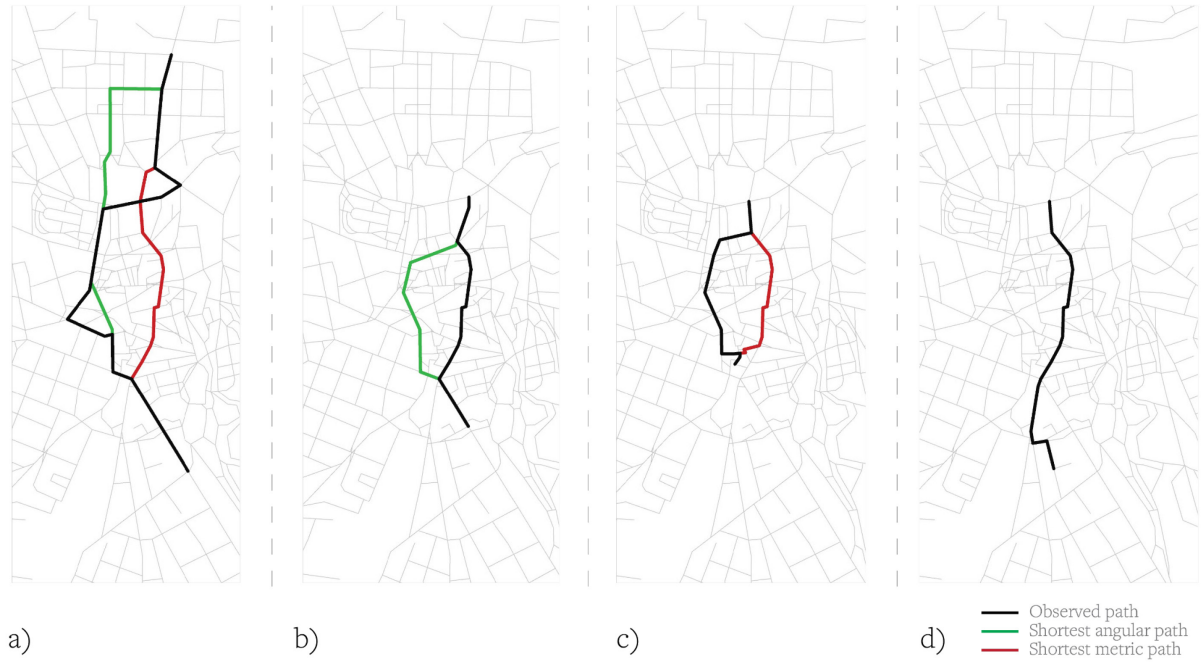


Figure 121. Examples of origins-destination pair with different accuracy of a cognitive and physical shortest path model. Black path = empirical path, Red path = shortest Euclidian path, Green path = shortest angular path.

Overall, we find that the cognitive model performs better with an accuracy of 85,6% compared to the physical distance-based model with 81,6% accuracy. When comparing the performance of both models on the level of an individual path (Figure 122), we observe that in some cases a) both models perform identically (points on the 45-degree diagonal line), b) the physical model is more accurate than the cognitive (points below the diagonal), c) cognitive model is more accurate than the physical (above the diagonal) and d) all three cognitive, physical and empirical paths are the same (point at [1,1] coordinates) as we find for the majority of cases.

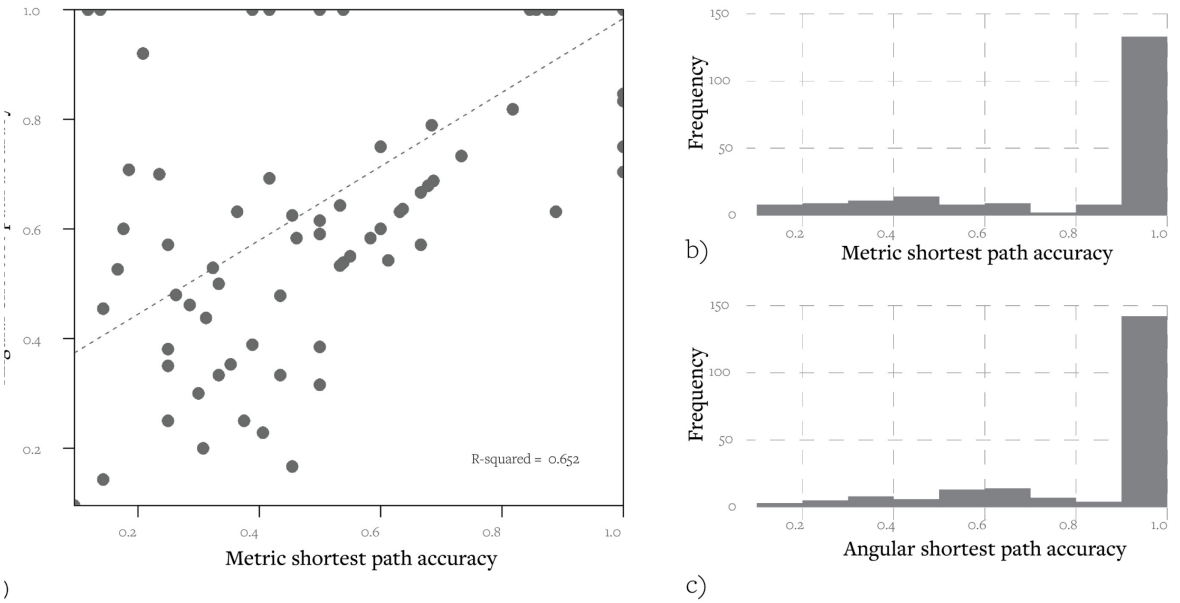


Figure 122. a) Scatterplot model accuracy for each observation. 45-degree diagonal line divides cases where either the physical or cognitive model performed better. Cases lying on the diagonal perform equally. b) histogram showing the performance distribution of the physical model. c) histogram showing the performance distribution of the cognitive model.

For a better understanding of the results, we examine a) the ability of cognitive shortest paths to minimize Euclidian distance and b) the ability of physical shortest paths to minimize angular deviation (see Figure 123). We found that on the one hand, the cognitive shortest paths are in the case of Weimar very close (highly correlated, $R^2 = 0.98$) to the minimal physical path (Figure 123a), but on the other hand, the physical shortest paths result in the large cognitive distance and the relationship between these variables is looser ($R^2 = 0.78$). This means that the cognitive model can find paths that are "easy" to navigate and not far away, whereby the physical model produces paths that are comparably far away as the ones from the cognitive model but significantly more difficult to navigate.

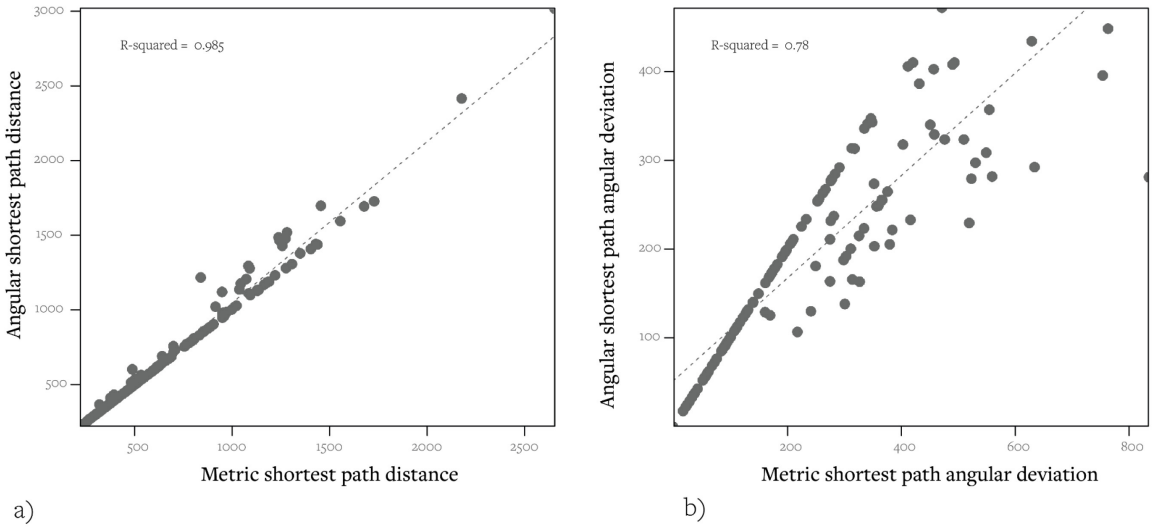


Figure 123. Comparison of the metric and cognitive shortest paths. a) Scatter plots comparing the cognitive and physical shortest paths in terms of Euclidian distance. b) Scatter plots comparing the cognitive and physical shortest paths in terms of angular deviation.

Summary

To summarize, we explored different computational models of path selection and tested them against empirical data. We conclude that in the case of Weimar, the cognitive shortest path model minimizing the angular deviation between origin and destination of travel was performing better than the metric alternative. Therefore, in the scope of this study, we adopt the minimal angular deviation as optimization criteria for path selection. Finally, we must note that since these results are not generalizable to other urban layouts and cannot be considered representative, further investigation is necessary.

Appendix 14 Travel Impedance Function

The integral part of a movement model adopted in this study is the travel impedance function expressing the willingness to travel as a function of distance. Traditionally, the simplest way to express how far a person is willing to travel is the Boolean function as used in the cumulative accessibility (Handy & Niemeier, 1997), closeness (Sabidussi, 1966), or betweenness centrality (Freeman, 1977). It is based on the assumption that the willingness to travel to a given destination is a dichotomous function of distance being either 1 or 0, based on the radius threshold. It means that up to a given distance, we are equally willing to travel, and beyond it, we do not travel at all.

This assumption has been empirically proven as unrealistic and replaced by continuous decay function commonly known under the term gravity-based impedance functions (Kwan, 1998; Vale & Pereira, 2016). Here, the willingness to travel continuously drops with the increasing distance, or in other words, the attractiveness of the destination rises as it gets closer. In this research, we employ the negative exponential function as a well-established type of gravitational function used to capture travel behavior (Equation (46).

$$f(d_{ij}) = \frac{1}{e^{\beta d_{ij}}} \tag{46}$$

The parameter defining the shape of the impedance function – the beta coefficient β is calibrated from empirical data capturing the actual travel frequency as a function of travel distance d_{ij}. It reflects the type of environment, travel mode, or travel destination considered in the movement model. In the following, we fit the negative exponential travel impedance function to capture the overall travel behavior of pedestrians in Germany and in the study area of Weimar.

The empirical data used to fit the beta coefficient comes from the MiD2017 study and comprises 186 673 travel records for the whole of Germany and 248 observations for the study area in Weimar. After filtering the journeys by the six travel activities adopted in this study, both sets get reduced to 60 002 observations for Germany and 77 observations for Weimar. On the one hand, the data set for Germany offers a high level of confidence due to the large sample size. On the other hand, the data set for Weimar offers local insights. For this reason, we evaluate both and compare the resulting travel impedance functions.

We start with extracting journey frequency table capturing the number of trips divided by distance range and travel activity. For this purpose, we calculate a cumulative trip frequency table where each column is the sum of all columns to its right side (Table 21).

Table 21. Cumulative journey frequency table for the whole of Germany (extracted from MiD2017). Rows are travel activities considered in the prediction model, and columns are travel distances. Each cell of the contingency table captures the number of trips toward the given activity below the given distance threshold.

	< 0.5km	< 1km	< 2km	< 5km	< 10km	< 20km	< 50km
Administrative	714.45	270.57	117.54	0	0	0	0
Education	15102.98	8704.07	3481.17	658.37	89.55	19.51	4.51
Gastronomy	4750.53	2531.13	1306.89	484.11	95.04	17.14	0
Healthcare	2135.22	1247.46	482.31	129.69	0	0	0
Shopping	12899.38	6019.24	2346.52	583.42	64.66	25.71	0
Work	12614.3	7637.37	3981.34	1511.39	373.75	163.63	58.63

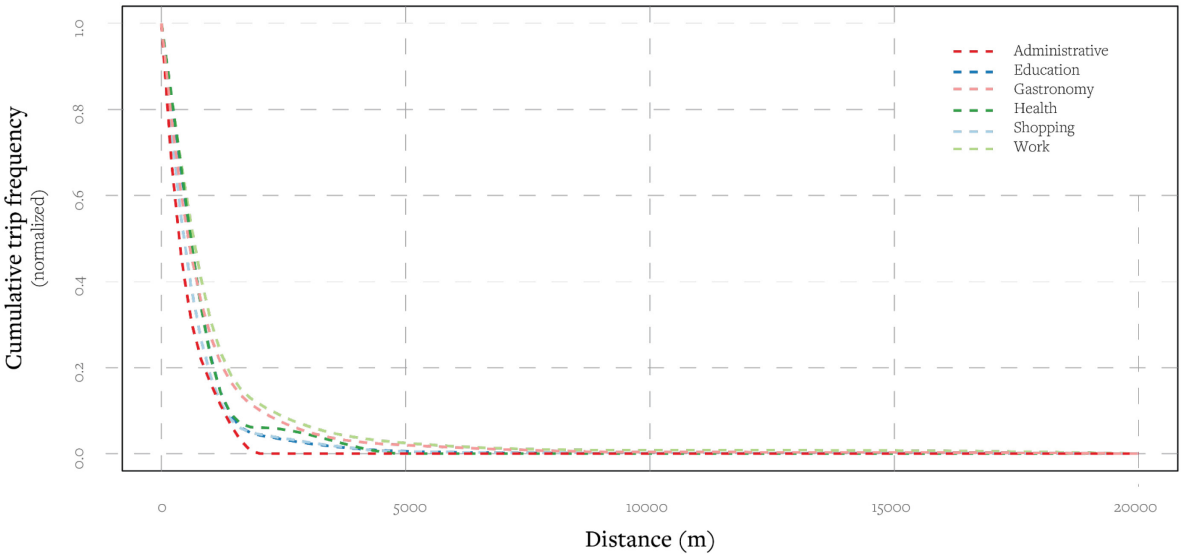


Figure 124. Cumulative journey frequency table for the whole of Germany (extracted from MiD2017). Curves are based on cubic spline interpolation of the original sample with df = 6 to n = 100 using method = "hyman".

As next, we test if the empirical travel impedances for each activity differ from each other or can be considered as following the same underlying distribution. This would simplify further model estimation procedure since a reduced number of different distributions means a reduced number of beta coefficients, which have to be estimated.

For this purpose, we test the difference in the impedance distribution via Kolmogorov–Smirnov test (KS test) and paired T-test to determine whether the mean difference between the two curves equals zero.

The KS test is a nonparametric test of the equality of continuous, one-dimensional probability distributions. It is used to compare the shape of distribution between two samples. The major advantage of the KS test is that no assumptions about the underlying distribution are necessary (it does not need to be normal). The null hypothesis is that both

samples are drawn from the same distribution. For both KS and T-test, we calculate the statistic for all 15 combinations (Equation (47) of two pairs from six considered activities.

$$C_r^n = \frac{n!}{r!\,(n-r)!} = \frac{7!}{2!\,(7-2)!} = 15 \tag{47}$$

We conclude that all 15 tests for both KS and Students T-test were not significant (confidence level = 0.95). This means that in all cases, we could not reject the null hypothesis stating that the travel impedance for all activity types comes from one underlying distribution. Consequently, we fit one joined travel impedance function and use it for modeling all six considered travel destination activities.

For this purpose, we join the rows of the travel frequency table by summing up the travel frequencies for each travel distance threshold (Table 22).

Table 22. Total travel frequency

	< 0.5km	< 1km	< 2km	< 5km	< 10km	< 20km	< 50km
Total travel frequency	49779.35	27528.45	12375.29	3673.88	670.52	234.56	63.14

Since the travel impedance function represents the willingness to travel and is standardized to range from 1 to 0, we must rescale the empirical travel frequencies to the same range before the fitting procedure. To match the travel impedance function, we adjust the distance range to start at zero distance. With growing distance, the willingness to travel modeled by exponential travel impedance function drops from 1 and asymptotically approaches 0 (Table 23).

Table 23. Normalized travel frequency

	>0km	>0.5km	>1km	>2km	>5km	>10km	>20km
Normalized travel frequency	1.000000	0.552441	0.247648	0.072627	0.012216	0.003447	0.000000

To fit the beta coefficient of the exponential travel impedance function, we calculate a sequence of 1000 impedance curves (from 0.0001 to 0.01 in 0.0001 steps) and search for the one best matching the empirical data. For each impedance curve, we calculated the goodness of fit measure as the mean squared distance between the empirical observations and the respective points on the curve (Figure 125). The best-fitting curve (i.e., the smallest mean square distance) is the one best capturing the empirical data. Afterward, we repeat the same procedure for the Weimar data.

We found that for the whole of Germany, the beta coefficient $\beta = 0.00124$ (Figure 125a) fits best the empirical data, while in Weimar, the best fit was found at $\beta = 0.00138$ (Figure 126a). In both cases, we observe that the beta coefficient is very sensitive in the range from 0 to the optimal fit, and afterward, even a large increase in the coefficient has only a marginal impact on its error.

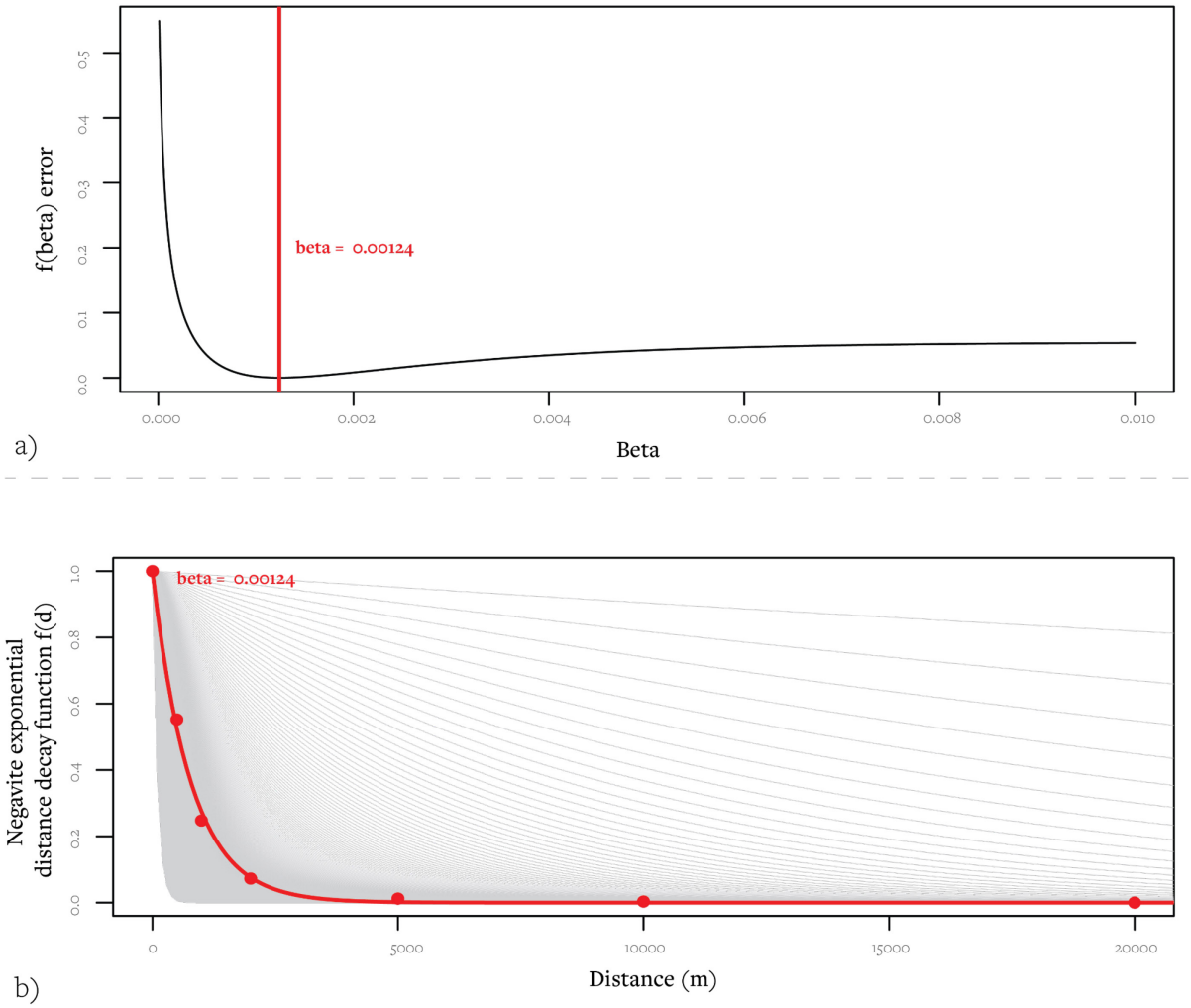


Figure 125. a) Relationship between impedance curve fitness (mean square distance) and the beta coefficient for the whole of Germany. The best fitness is the lowest point on the curve (i.e., with the smallest distance to the empirical data). For the pedestrian frequencies capturing the whole of Germany, the best fit was found for the beta coefficient 0.00124. b) Beta coefficient fitting to empirical data for pedestrian travel impedance in whole Germany. The best fit was found for the beta coefficient of 0.00124.

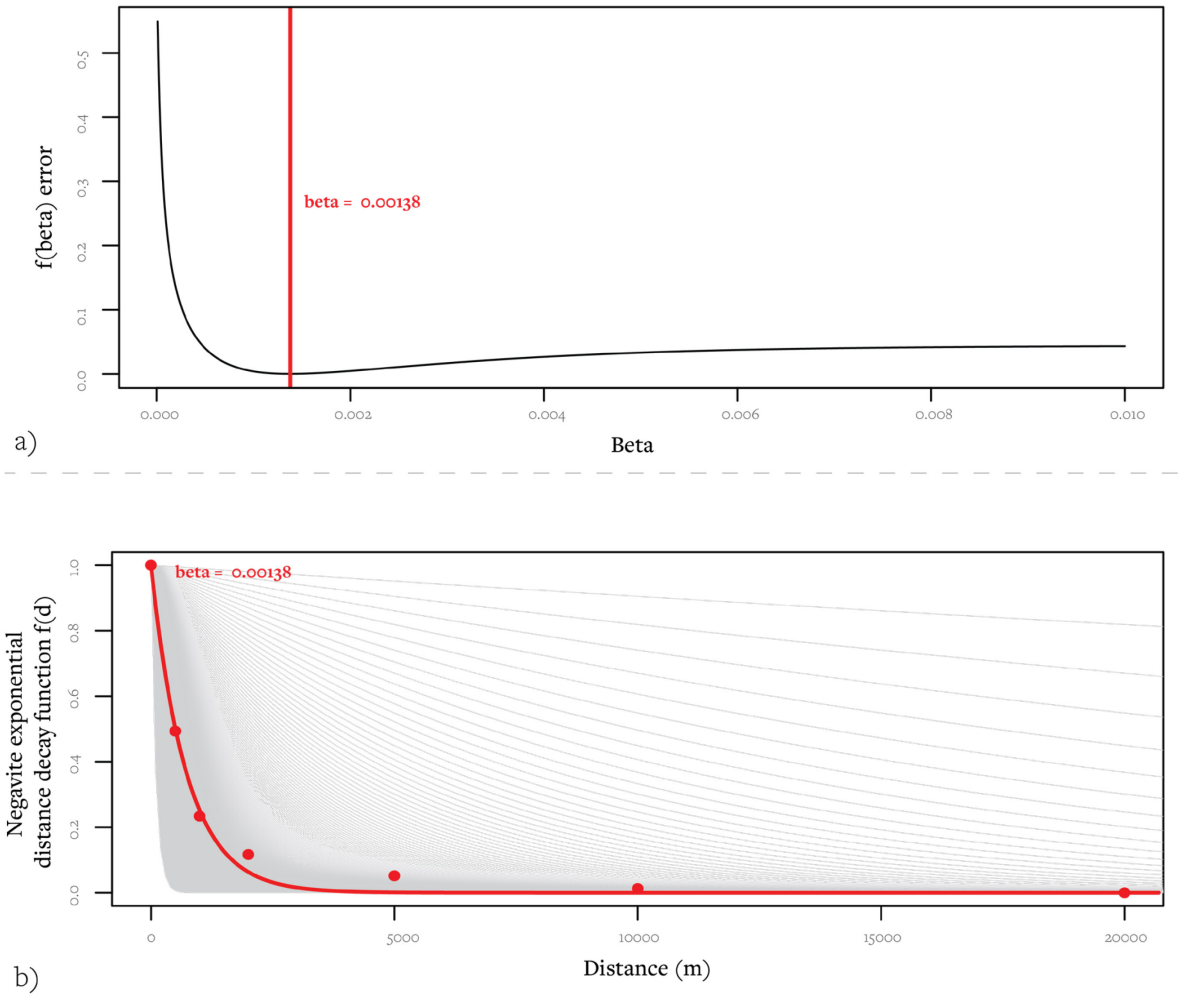


Figure 126. a) Relationship between impedance curve fitness (mean square distance) and the beta coefficient for Weimar. The best fitness is the lowest point on the curve (i.e., with the smallest distance to the empirical data). For the pedestrian frequencies capturing the pedestrian impedance in Weimar, the best fit was found for the beta coefficient 0.00138. b) beta coefficient fitting to empirical data for pedestrian travel impedance in Weimar.

Finally, we compare the fitted exponential travel impedance functions for a) Weimar, b) Germany, and c) transportation planning literature (Handy & Niemeier, 1997).

The commonly used travel impedance function in transportation planning literature was fitted by Handy & Niemeier in 1997. It is based on shopping travel diaries from 1980 of the Metropolitan Transportation Commission, 101 Eighth Street, Oakland, CA 94607. The respective beta coefficient in the temporal unit was 0.1813, which corresponds to 0.0018 in km. The second beta coefficient fitted for the whole of Germany was 0.00124, and the third beta coefficient was 0.00138. This means that both travel impedance curves fitted for Germany and Weimar are less steep than the established curve fitted in the North American context (Figure 127). Additionally, we conclude that pedestrians at the national level are more willing to overcame larger distances than in the case of Weimar even though both beta

coefficients are relatively similar. The relationship between the German and North American pedestrian travel impedance functions reflects the notoriously less walkable and more car-oriented environment of American cities.

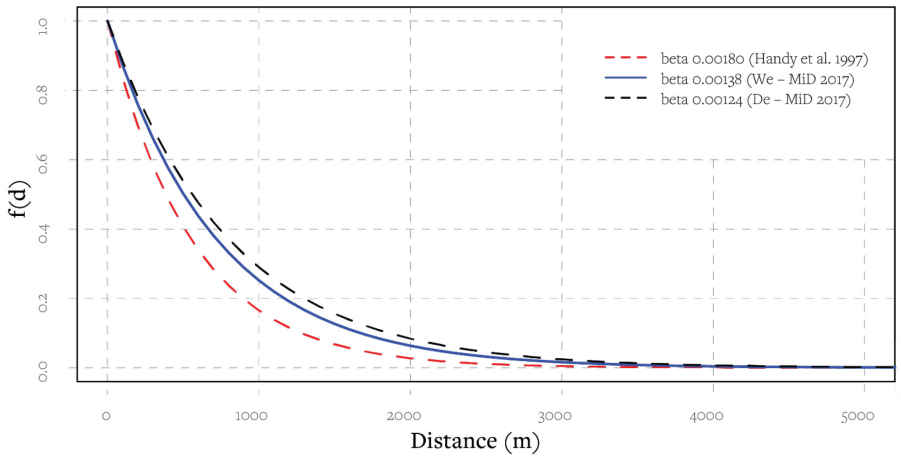


Figure 127. Calibrated negative exponential distance decay function with beta coefficient $\beta = 0.00138$. The function for Weimar (blue curve) is compared with the average for the whole of Germany (black curve) and reference function fitted in 1997 by Handy et al.

Summary

We conclude that the results of the impedance curve fitting procedures are in accordance with the existing literature sources and confirm the expected local differences in the impact of distance on the walking behavior. For the purpose of modeling the pedestrian movement in Weimar, we adopt the exponential travel impedance function with the beta coefficient 0.00138.

Appendix 15 DecodingSpaces Toolbox for Grasshopper

DeCodingSpaces Toolbox for Grasshopper is a collection of analytical and generative components for algorithmic architectural and urban planning. The toolbox is free software released by the Computational Planning Group (CPlan) and is a result of a long-term collaboration between academic institutions and praxis partners across the globe with the common goal to increase the efficiency and quality of architecture and urban planning.

The authors of this study are founding members and active contributors to the toolbox, with a large part of the tools (e.g., street network analysis toolbox) being specifically developed for purposes of this research. For all computational analyses presented in this study, we used the version 2019.07 accessed and published in July 2019 from https://toolbox.decodingspaces.net/.

Appendix 16 Movement Simulation Engine Distance Calculation Artifacts

The implementation of the DeCodingSpaces components for Grasshopper (v09.2019) used throughout this research two methods for computing path lengths. The first method (Method A) considers the start and endpoint of each journey as the midpoint of the selected street segment (Figure 128). The second approach to path length calculation assumes the longest possible path between the segments at the origin and destination of movement (Figure 128).

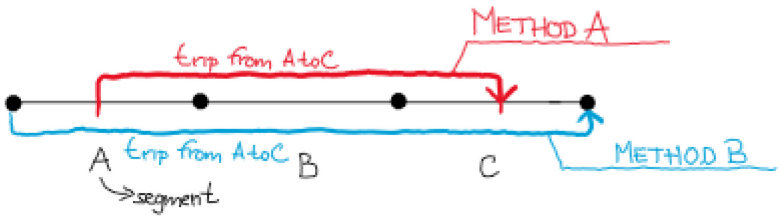


Figure 128. Illustration of two different trip length measurement methods.

We must note that for the calculation of the *From-Volume* the method A is used, while for the calculation of the *Through-Volume,* method B is employed. The actual difference between both methods affects only the start and end segment of the path. This means that the longer the path (in the number of segments), the smaller the difference and both methods. For future research, only one method should be used for the calculation of all movement characteristics to avoid inconsistencies.

Appendix 17 Filter Model

Multiple binomial logistic regression (further only logistic regression) is a statistical model belonging to the family of generalized linear models. It is used to model the probability of two discrete outcomes (e.g., "Success" vs. "Failure," or in our case, "Positive" for activity level higher than zero vs. "Negative" for zero activity level). It takes multiple continuous explanatory variables (i.e., exogenous and endogenous movement), fit sigmoid function, and return the probability of belonging to a certain class in the range from 0 to 1. Formally, logistic regression uses the logit link function. This is the logarithm of the odds $ln\left(\frac{p}{1-p}\right)$ with p standing for the probability of success. The advantage and main purpose of the logit function is to map of probability values constrained by a range of (0,1) to range $(-\infty,+\infty)$ so they can be fitted by linear model:

$$ln\left(\frac{p_{Activity}}{1-p_{Activity}}\right) = \alpha_0 + \alpha_{en} * M_{en} + \alpha_{ex} * M_{ex} + \epsilon \tag{48}$$

This model can be then used to assess the probability level for any value of the explanatory variable:

$$P(Activity = 1|M_{ex}, M_{en}) = \frac{exp(\alpha_0 + \alpha_{en} * M_{en} + \alpha_{ex} * M_{en}}{1 + exp(\alpha_0 + \alpha_{en} * M_{en} + \alpha_{ex} * M_{ex}} \tag{49}$$

By plugging the value of the pedestrian *Through-Frequency* movement, the model returns the probability of non-zero activity level. In other words, what is the probability of finding any shopping, gastronomy, education, work, administrative, or healthcare-related activities. It is important to realize that the logistic model does not tell how much, but rather if any activity can be found.

The logistic regression model can be used as a classifier by choosing the probability cutoff value and diving inputs below and above this value in two classes. The cutoff choice is essential for the classifier sensitivity and specificity and the overall performance of the model. The terminology is coming from medical research and describes the performance of the diagnosis test. Sensitivity represents the percentage of sick individuals being correctly diagnosed by the test as sick, and the specificity captures the percentage of healthy individuals being correctly diagnosed as healthy. Differentiating between both categories instead of counting the overall number of misclassifications is crucial when working with an unbalanced data set. In the case of a rare illness which affects only 0.1% of the population, we can achieve 99,9% overall accuracy of the test just by classifying every person as healthy. However, such a test is as good as no test, which can be revealed by looking at its sensitivity. This is of special interest if the ability to identify positives (sick individuals) is more

important than the overall accuracy. By adjusting the cutoff value, we can steer the trade-off between sensitivity, specificity, and overall misclassifications to any direction required. Coming back to our context, we conclude that most of the activities considered in this study are highly unbalanced. With a large portion of the street segments with no other activity other than living, we face similar difficulties as with the "rare illness" example. Therefore, we prioritize the sensitivity over specificity to calibrate the cutoff level for each activity type to correctly classify at least 90% of street segments with activity levels above zero.

In the following, we summarize the attributes and performance of all six logistic regression models. We formally define all terms and evaluation criteria and accompany the numerical results with graphical plots.

Logistic Model Summary

We reduce the interpretation of the coefficients to their sign as their amplitude represents the change in log odd ratio of logarithmically transformed explanatory variables on a logarithmically transformed dependent variable. As a result, any attempt to interpret the magnitude of the coefficient will be a mathematical exercise with little practical implications.

We found that in four out of six logistic regression models, both movement components were highly significant predictors (Table 24). Only in the case of administrative activities, the exogenous movement was not significant, while for the educational activities, the endogenous movement was below the significance level (p-value < 0.05).

Table 24. Logistic regression model coefficients

	Estimate	Std.	Error	z	p-value	Pr(>|z|)
Administrative	(Intercept)	-7.1803	0.9754	-7.361	1.82E-13	***
Administrative	movement.exo	-0.4945	0.3226	-1.533	0.12533	
Administrative	movement.endo	0.8505	0.2948	2.885	0.00392	**
Education	(Intercept)	-4.75381	0.56054	-8.481	< 2e-16	***
Education	movement.exo	0.41227	0.11886	3.468	0.000523	***
Education	movement.endo	-0.10347	0.07255	-1.426	0.153799	
Gastronomy	(Intercept)	-8.3685	0.8628	-9.699	< 2e-16	***
Gastronomy	movement.exo	-1.1397	0.264	-4.317	1.58E-05	***
Gastronomy	movement.endo	1.5774	0.2541	6.208	5.36E-10	***
Health	(Intercept)	-8.9844	1.1237	-7.995	1.29E-15	***
Health	movement.exo	-1.6463	0.3371	-4.883	1.04E-06	***
Health	movement.endo	1.9756	0.3266	6.048	1.46E-09	***
Shopping	(Intercept)	-17.1875	2.2861	-7.518	5.55E-14	***
Shopping	movement.exo	-3.1695	0.4956	-6.395	1.60E-10	***
Shopping	movement.endo	3.9691	0.5486	7.235	4.67E-13	***
Work	(Intercept)	-1.08886	0.24728	-4.403	1.07E-05	***
Work	movement.exo	1.35576	0.12005	11.293	< 2e-16	***
Work	movement.endo	-0.96607	0.09752	-9.906	< 2e-16	***

Model Performance

We introduce and discuss different aspects of the logistic model performance. Based on the choice of cutoff value and purpose of the model, different characteristics might be relevant. We evaluate the robustness of the model in relation to the choice of the cutoff value, identify optimal cutoff value under which we measure the sensitivity, specificity, and misclassification error.

Receiver Operating Characteristics Curve

Receiver Operating Characteristics Curve traces the percentage of true positives accurately predicted by a given logit model as the prediction probability cutoff is lowered from 1 to 0. For a good model, as the cutoff is lowered, it should mark more of actual 1's as positives and lesser of actual 0's as positives. This means, the curve should rise steeply, indicating that the True Positive Rate (TPR) (Y-Axis) increases faster than the False Positive Rater (FPR) (X-Axis) as the cut-off score decreases. In effect, the greater the area under the ROC curve (AUROC), the better the predictive ability of the model.

Based on the AUROC model performance characteristic we found the shopping, gastronomy and health activity logistic model to perform well for the whole range of possible cut-off values (Table 25). For the remaining activities we must choose between high sensitivity and specificity of the model since it is not possible to achieve both at the same time.

Table 25. Area under Receiver Operating Characteristics Curve

	Administrative	Education	Gastronomy	Health	Shopping	Work
AUROC	0.73	0.65	0.81	0.84	0.91	0.75

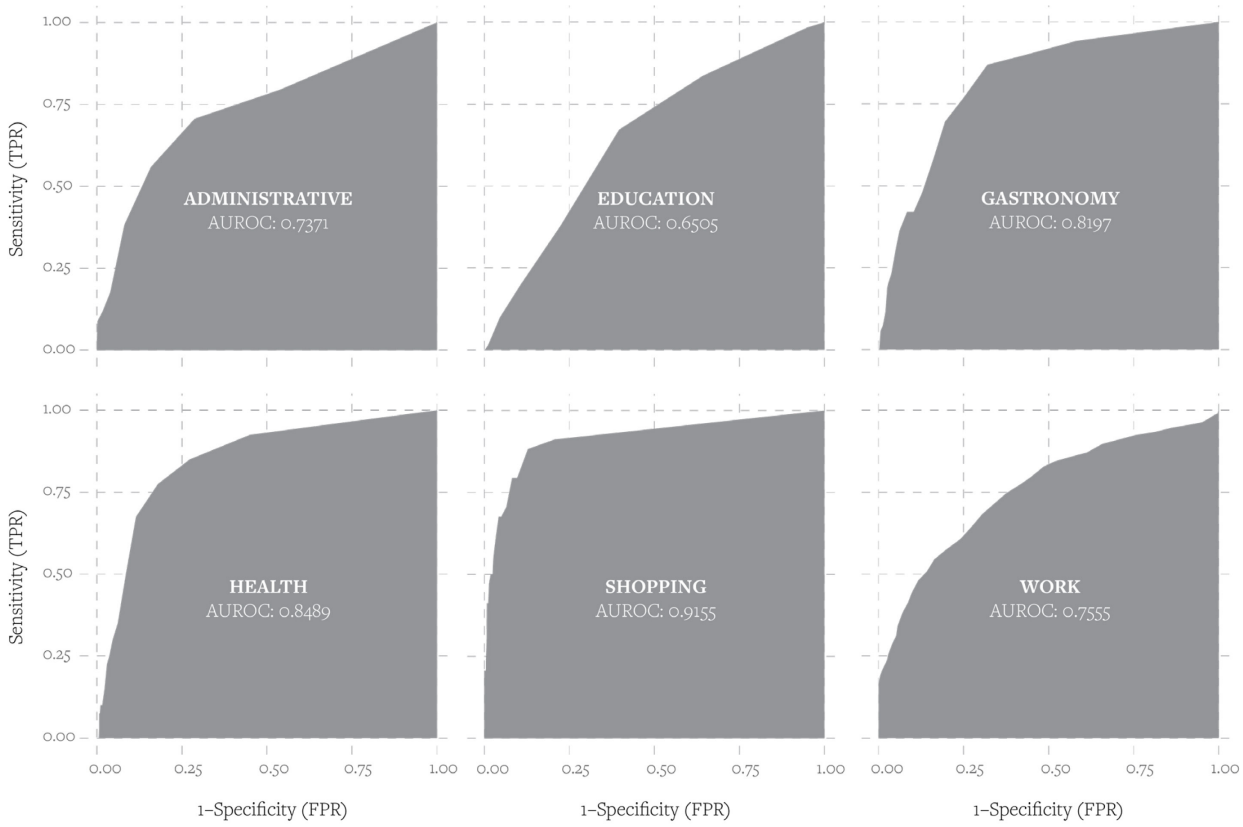


Figure 129. Receiver Operating Characteristics Curve for logistic models by activity type.

Cut-off Value

The optimal cut-off value for each logistic regression model (Table 26) was determined by maximizing the accuracy (i.e., lowest misclassification error) while keeping the sensitivity about a 90% level. By doing so, we can guarantee that at least 90% of all streets with non-zero work activity levels will be classified correctly. We conclude that the highly unbalanced dataset (i.e., more zero-activity street than non-zero activity streets) results in low cut-off values. For most activities except work, the cut-off is below 0.05. In other words, if the probability of any street containing activity is equal or higher than 5%, we classify such street as positive (i.e., we expect a non-zero activity level). Since the work activities are the most balanced activity type in the dataset (565 streets with non-zero activity levels and 482 zero-activity level streets), the cut-off value of 0.34 is closer to the 50% probability mark.

Table 26. Cut-off values for each activity type logistic regression model.

	Administrative	Education	Gastronomy	Health	Shopping	Work
Cut-off	0.01	0.036	0.05	0.02	0.029	0.34

Misclassification Error

Misclassification error of logistic regression is the percentage mismatch of predicted vs. actuals, irrespective of 1's or 0's. It is the ratio of trues to false. The lower the misclassification error, the better is the overall model performance for the given cut-off value.

$$Misclassification\ erro = \frac{truePositive + trueNegative}{falsePositive + falseNegative} \tag{50}$$

We found the best overall accuracy for the shopping activity model, with 83% of correct matches (Table 27). This is followed by work, gastronomy, and health activity ranging from 56% to 66% classification accuracy. The worst performing models are for education and administrative activity, achieving only 33% and 29% correct classifications.

Table 27. Misclassification error for each activity type logistic regression model.

	Administrative	Education	Gastronomy	Health	Shopping	Work
Misclassification error	71%	67%	34%	44%	17%	36%

Sensitivity & Specificity

Sensitivity (or True Positive Rate) is the percentage of 1's (actuals) correctly predicted by the model, while specificity is the percentage of 0's (actuals) correctly predicted. Specificity can also be calculated as 1 - False Positive Rate.

$$Sensitivity = \frac{\#\ Predicted\ True\ Positives}{\#\ Actual\ Positives} \tag{51}$$

$$Specificity = \frac{\#\ Predicted\ True\ Negatives}{\#\ Actual\ Negatives} \tag{52}$$

Since the cut-off value was specifically selected to favor sensitivity, for all six models, we observe similar sensitivity levels above 90%. However, the logistic regression models differ significantly when it comes to specificity (Table 28). The highest specificity level is observed in the case of the shopping activity model being able to classify 83% of all zero-activity observations correctly. On the contrary, the lowest performance was measured for the administrative activities with the model specificity of only 27%.

Table 28. Sensitivity and specificity for each activity type logistic regression model.

	Administrative	Education	Gastronomy	Health	Shopping	Work
Sensitivity	94%	90%	90%	92%	90%	90%
Specificity	27%	30%	64%	55%	83%	34%

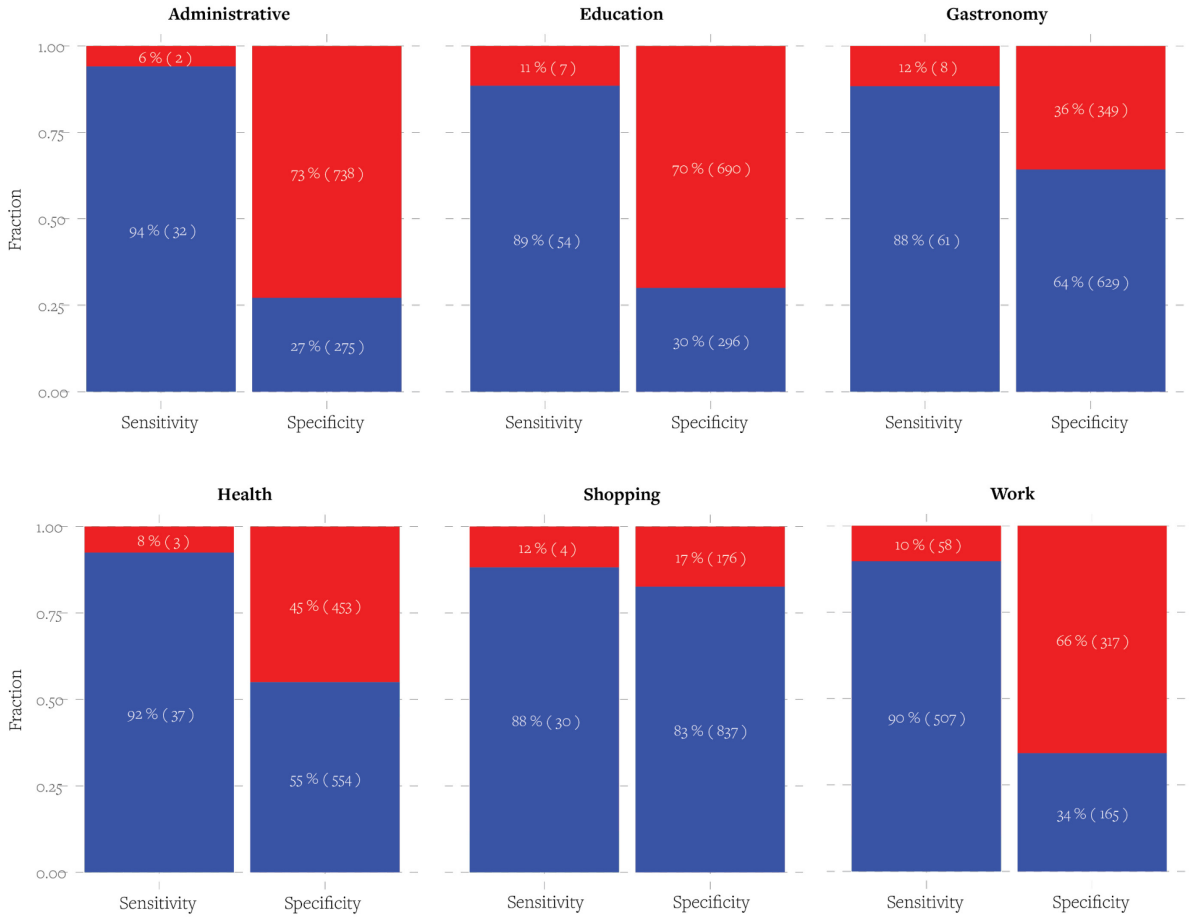


Figure 130. Stacked bar plot showing the sensitivity and specificity of each logistic regression model by activity type. Correctly classified observations are marked blue, misclassified observations are marked red.

Logistic Curves, Classified Positives & Negatives

We visualize the logistic curve as fitted by the model for both explanatory variables. The displayed curves for each variable are constructed by holding the other variable constant at its mean value. The overall probability and resulting classification are results of the interaction between both curves. These figures confirm what was already visible from the fitted regression coefficients - namely, the direction in which endogenous and exogenous movement influence the presence or absence of given activity.

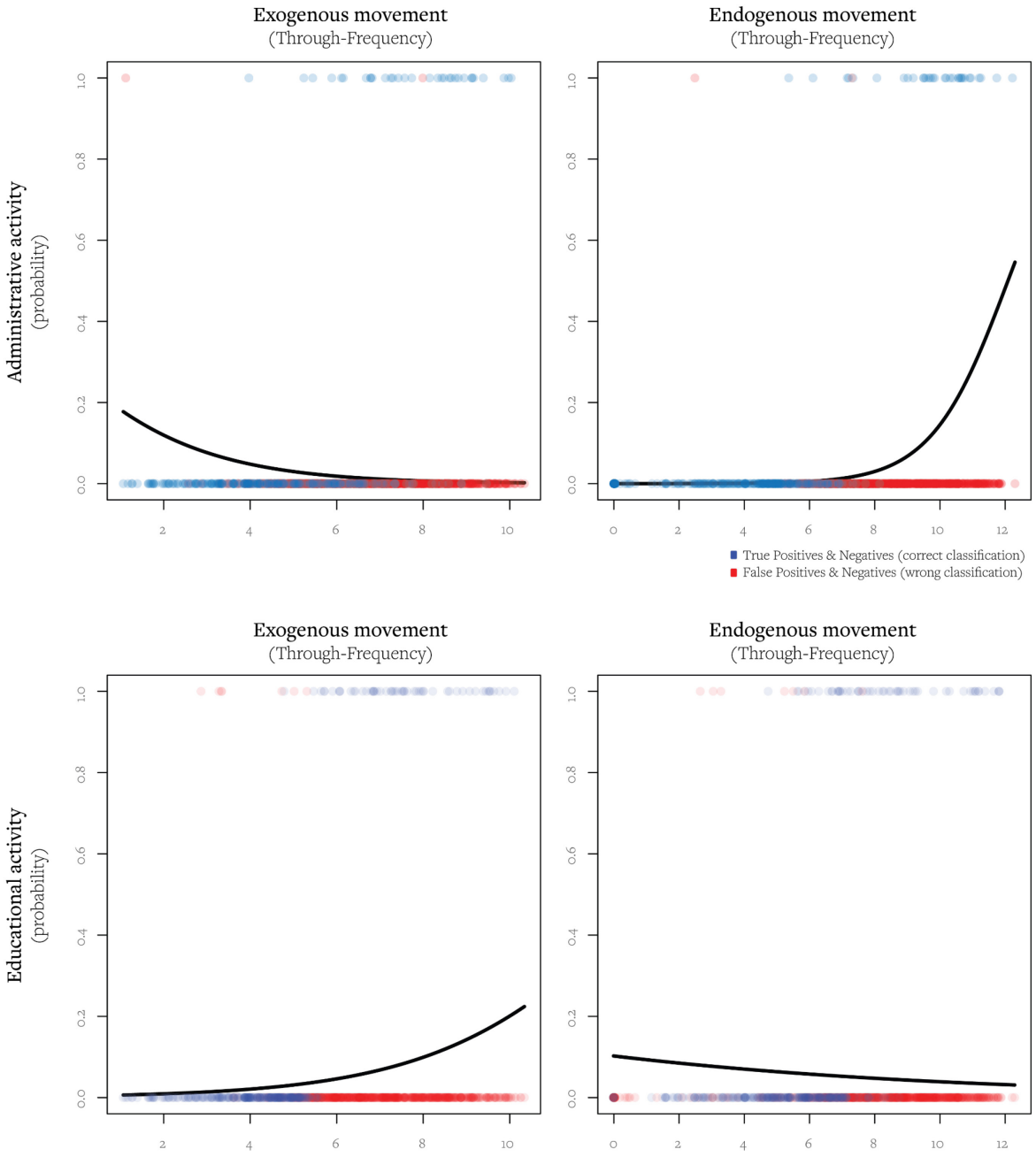


Figure 131. Logistic curve for Administrative and Educational activity type as fitted by the filter model. The displayed curves for each variable (i.e. endogenous and exogenous movement) are constructed by holding the other variable constant at its mean value. Correctly classified observations are marked blue, misclassified observations are marked red.

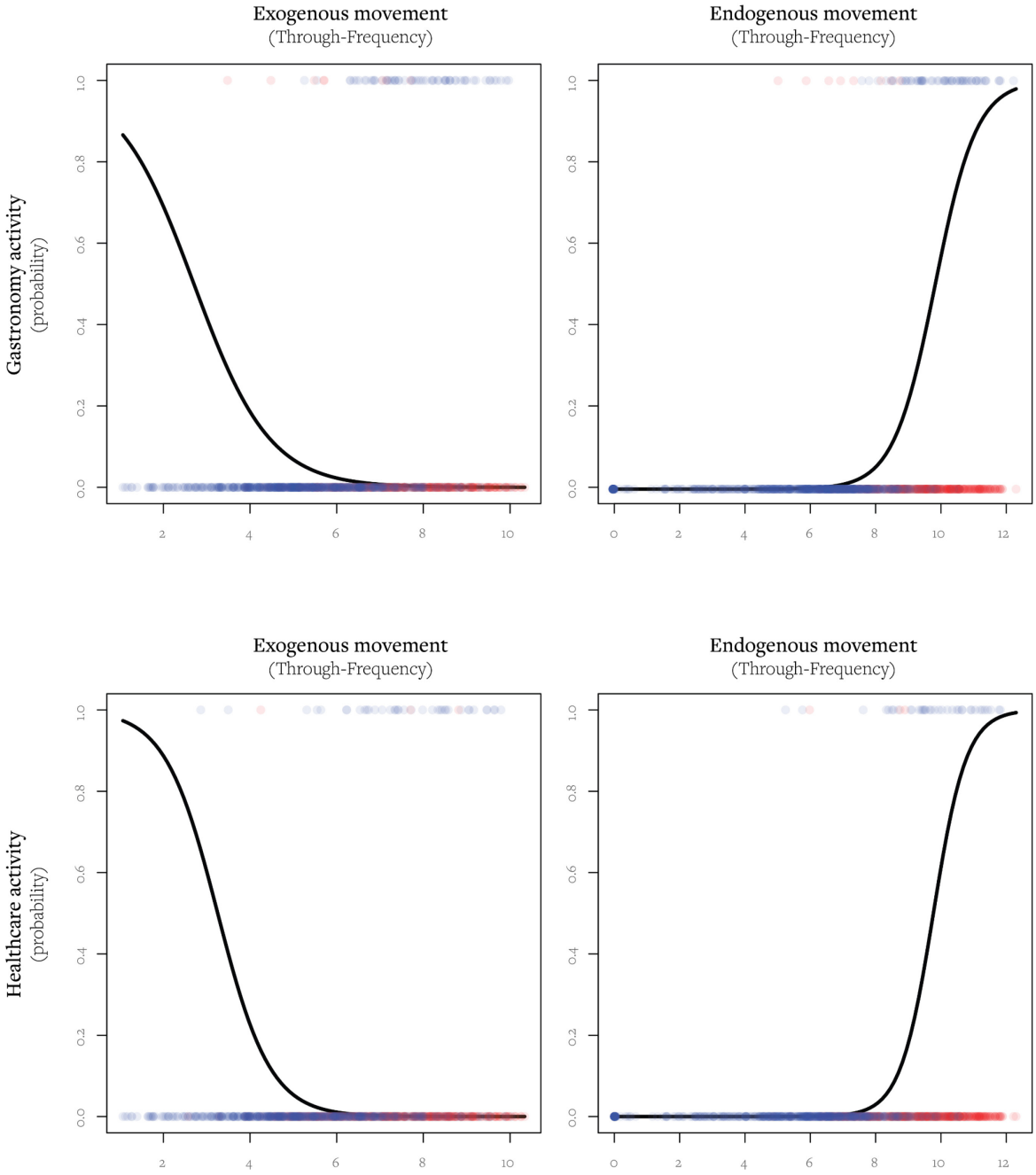


Figure 132. Logistic curve for Gastronomy and Health activity type as fitted by the filter model. The displayed curves for each variable (i.e. endogenous and exogenous movement) are constructed by holding the other variable constant at its mean value. Correctly classified observations are marked blue, misclassified observations are marked red.

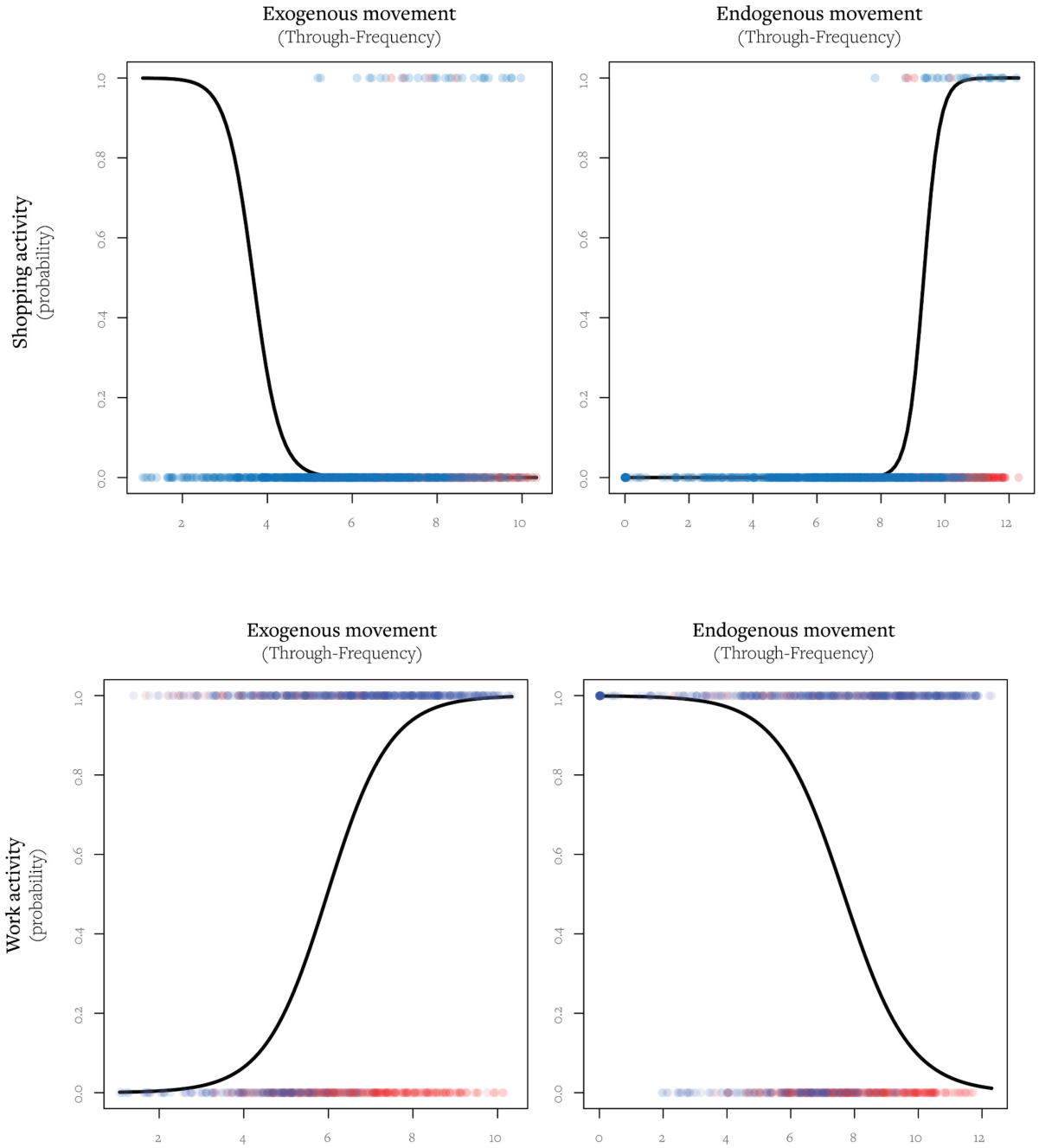


Figure 133. Logistic curve for Shopping and Work activity type as fitted by the filter model. The displayed curves for each variable (i.e. endogenous and exogenous movement) are constructed by holding the other variable constant at its mean value. Correctly classified observations are marked blue, misclassified observations are marked red.

Summary

We found that the logistic regression model with exogenous and endogenous movement as explanatory and shopping activity as the dependent variable was performing best across all criteria. It correctly classified 90% of non-zero shopping activity street segments and filter out 83% of street segments without any shopping activity. The corresponding model for gastronomy and health activity was able to filter out 64% and 55% of respective zero-activity street segments (see specificity in Table 29). For all three activity types, both movement components were significant predictors of street segments with non-zero activity levels. For the remaining work, education, and administrative activities, the filter performance drops sharply to 34%, 30%, and 27% sensitivity. This means that a large portion of zero-activity street segments remains in the dataset.

Table 29. Summary table of the *Filter* (multiple logistic regression model) by activity type.

	Administrative	Education	Gastronomy	Health	Shopping	Work
Exogenous movement	/	+	-	-	-	+
Endogenous movement	+	/	+	+	+	-
AUROC	0.73	0.65	0.81	0.84	0.91	0.75
Cut-off	0.01	0.036	0.05	0.02	0.029	0.34
Sensitivity	94%	90%	90%	92%	90%	90%
Specificity	27%	30%	64%	55%	83%	34%
Misclassification error	71%	67%	34%	44%	17%	36%

Appendix 18 Amplifier Model

Motivation is to model the effect of movement flow on the activity levels while considering spatial interactions between the activities. It is the later - spatial interaction between the activities what presents major challenges as the presence of spatial dependence violates some of the linear regression assumptions (Anselin & Rey, 2014):

- Observations are independent
- Error terms are not correlated with explanatory variables
- Error terms have constant variance
- Omitted variables cannot be correlated with any explanatory variable include in the model

However, if we ignore spatial dependence, we ignore substantive spatial interaction driving the process behind the observed pattern. This induces that the ordinary least squares estimator (OLS) is a) biased and b) inconsistent. The former means that the expected estimated model parameters are significantly different than the true population parameters, and the latter implies that even with increasing sample size, the parameters do not get corrected. To summarize, by ignoring spatial interaction of any kind, we get wrong estimates of the model parameters characterizing the effect of individual explanatory variables as well as wrong significance levels of each predictor.

As a result, if spatial dependence is present, and this is what we assume based on the literature review on the urban economy, we cannot ignore it, and at the same time, we cannot model it via standard linear regression. For this reason, we turn to a different class of regression models - the Mixed Regressive-Spatial Auto-regressive linear model (SARLM). This model makes it possible to eliminate the endogeneity bias caused by spatial dependence in activity intensity and avoid the omitting variable bias caused by ignoring the spatial dependence.

SARLM General Principles and Model Specification

We briefly summarize the core principles of SARLM to shed light on the model diagnostic and specification test. As it is the case in any type of modeling, in statistics, we abstract and simplify complex reality to models that do not aim to be true but rather useful. However, they are useful only under specific conditions or assumptions which always must be tested. Therefore, we do not simply run SARLM model and report its results; we test the specification of the model in the first place.

We define the SARLM model and its relationship to LM model:

$$LM: y = \alpha X + \mu \tag{53}$$

$$SARLM: y = \rho Wy + \alpha X + \mu \tag{54}$$

In our case, this translates to:

$$activity = \rho Wactivity + \alpha Movement + \mu \tag{55}$$

where $activity$ is a vector of activity levels at different locations, W stands for spatial weights matrix, $Wactivity$ is spatial lag, ρ is a spatial auto-regressive coefficient, and $Movement$ is a matrix of dependent variables (endogenous and exogenous movement), α is a vector of coefficients for each dependent variable and μ is the error term.

In general, the best model is the simplest one, which serves the purpose. Every extension of the model takes more uncertainty and risk of bias, which always must be weighed against the possible gains in precision. Therefore, we start with a simple linear regression model and test for spatial dependence in activities. If and only if we detect spatial interaction in activity allocation, we extend the LM to SARLM account for it.

The regression model coefficients can be estimated through several different methods, such as two-stage least squares, maximum likelihood, or general methods of moments. We have chosen the maximum likelihood as it is most efficient among all consistent estimators and thus works best with small samples[43] (Anselin, 1988).

On the most abstract level, we can say that the spatial autoregressive model is about modeling interaction without interaction data. In other words, we do not observe the interaction itself (i.e., the process); we see only its result (i.e., the pattern) as the data we have is only cross-sectional. In essence, we try to solve n^n problem, with only n observations. As it is common practice in statistics, when information is needed but not available, we replace it with assumptions. In the following, we discuss the core assumptions behind the SARLM model and their implications.

Spatial Weights

To simplify the interaction problem, we assume that only spatially related observations interact. In other words, we superimpose structure in the form of a weights matrix defining how observations are spatially related. There are multiple ways how this can be achieved, each with substantial consequences for the model characteristics. It is important to emphasize that this structure is not coming from the data itself - it is not empirical, but it is defined by us - it is theoretical. The spatial weight matrix (W) is what ties together the whole model and helps correct for the major violation of linear regression assumption about

[43] For some activities (e.g. gastronomy and shopping) the filtered sample size is less than 50 observations.

the independence of observation. We assume that these are not independent, and their interdependence is captured in the W. In general, this matrix captures the relationship between the observations, which is represented either in the form of contiguity or distance-based weighting (Figure 134).

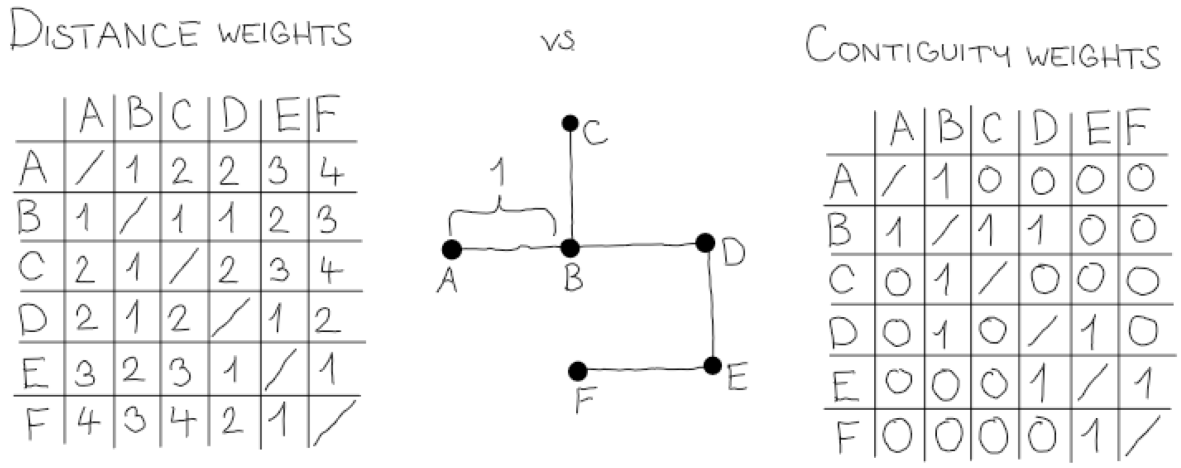


Figure 134. Exemplary graph with distance and contiguity weights matrix.

The decision of how we define neighborhood influences the way how the economic and social spillover generated at one location propagate through the system. In the distance-base weight matrix, we recognize that everything is related to everything else, with closer neighbors being more related than the distant ones. On the contrary, the contiguity weights operate on the basis of a hard threshold considering all observation being either neighbor or non-neighbor. The threshold defines the order of neighbors being considered (e.g., second-order means that we consider the neighbors of the neighbors); however, without the ability to differentiate between them. In the case of the contiguity weights, all neighbors, regardless of their distance, are considered equal [44].

From a computational point of view, the contiguity weights are far more efficient when a) constructing the weight matrix, b) storing it, and, most importantly, c) when estimating the regression model coefficients. In all three cases, the advantage of contiguity weight over distance base weighting can be traced back to the fact that most of the matrix is filled with zero values

Even though from a behavioral point of view, the distance-based weight matrix might better fit our purpose, we conclude that due to its computational limitations, it is at current times

[44] It has to be noted, that the variance - co-variance structure of SARLM process is global. It means that the spillovers are propagating beyond their direct neighbors even in the case of contiguity weight matrix. This structural global property of SARLM to some degree compensates the otherwise local nature of the contiguity weight matrix.

not practically feasible. In this study, we utilize the spdep [45] package for R [46] to estimate the SARLM model, which supports due to already discussed reasons only contiguity weights. Its current implementation does not allow for isolates in the weight matrix specification. As a result, all observations with no direct neighbors must be excluded from the regression. This is significantly reducing the data set available to estimate the regression model, as many observations remain isolated after the first step of filtering zero-activity segments. Consequently, it is not meaningful to analyze spatial dependency in data whit little to no spatially related observations. To mitigate the negative effect of contiguity weights, we extend the definition of neighbors to second-order relationships.

Instrumental Variables

The apparent problem of the SARLM specification is that the dependent variable y is on both sides of the equation. Why is this a problem? It is simply not possible to tell them apart. We want to know which portion of y is influenced by other y and which is caused by other factors. To untie this Gordian knot, we use a well-established concept known as instrumental variables (Wright, 1928) - additional variables that isolates the problematic portion in y and make the estimation possible. As a result, in SARLM, the spatial dependent autoregressive variable y is not a function of its neighboring values as one might expect by reading the model specification. It is rather a function of the characteristics of its neighbors as these are used to construct the instrument. This means that the activity level at a given location is not based on the activity levels at other locations but rather on the movement that characterizes the activities at neighboring locations. This might be more visible when looking at the reduced form of the model specification, which directly implies the instruments used to account for the self-dependency.

$$y = (I - \rho W)^{-1}\alpha X + (I - \rho W)^{-1}\mu \tag{56}$$

Equilibrium

SARLM is to be interpreted as the equilibrium outcome of the spatial interaction process, a spatial reaction function (Brueckner, 2003). It is assumed that the simultaneous interaction between y and its neighbors Wy eventually arrive at a point of equilibrium, which is the point when the empirical data was collected. This is a crucial assumption as many systems might never achieve this point. However, in the case interaction between activities and movement, the simulation experiments conducted by authors of this study show that these systems tend to find such a point of equilibrium (Bielik et al., 2019). Nevertheless, it remains unclear if any particular urban system already reached the equilibrium stage or is still on its way there. For this reason, we have deliberately chosen

[45] spdep version 1.1-3

[46] R version 3.6.2

our study site in the historical medium-size city of Weimar with a stable population and a large portion of its area under cultural heritage protection.

Spatial Asymptotics

It is assumed that as the sample grows, we have more information and thus are better at estimating the model coefficients. In order to make this *spatial asymptotics* assumption valid, the so-called *regularity conditions* have to be fulfilled. These are stating that regardless of where we look, we observe the same process. In reality, processes are often changing from one place to another, and if we do not control for this change (e.g., with additional explanatory variables), new data might not improve the model. Additionally, we assume that the degree of spatial heterogeneity is restricted and that spatial dependence occurs only in a specific range. This is crucial since if everything is strongly related to everything else, then new data does not bring much new information. The spatial asymptotics assumption gets often violated in case of large-scale development where everything is more or less the same. In such a case, it is obvious why adding new identical observations to the data set where everything is the same does not help. On the contrary, in statistics, everything is about variation. We argue that due to the historically grown and typologically rich small-scale urban structure of Weimar, the spatial asymptotics assumption is fulfilled.

Constrained Model - Linear Regression Model

We estimate simple multivariate linear regression with activity intensity per street segment as a dependent variable and exogenous and endogenous pedestrian *Through-Volume* movement as two explanatory variables. The observations used to estimate the model are based on the output of the *Filter* (i.e., multiple logistic regression). By design, they account for at least 90% of the non-zero activities. Identically to the filtering, we predict each individual activity separately and use the aggregated movement (i.e., generated by all activities together) as an explanatory variable.

$$activity = \alpha_0 + \alpha_1 movement_{exogenous} + \alpha_2 movement_{endogenous} + \mu \qquad (57)$$

Model Performance

As the measure of fit of the linear model, we usually report the R-squared and adjusted R-squared. This can be interpreted as the variance explained by the model and has the neat property ranging from 0 to 1. The adjusted version of R-squared is additionally to the variance explained also adjusts for predictors that are not significant in a regression. Overall, R-squared is easily understandable and comparable across different models.

Unfortunately, as soon as we deal with other types of regression models such as SARLM, the calculation of R-squared is no longer possible. Even though there are different versions of pseudo-R-square measures, these are a) not comparable across different models, and b)

do not describe the variance explained. They might be useful when testing different specifications of the same model but no more than that.

Since the linear regression model is only a base model used to justify the necessity of the spatial autoregressive model, we need a measure of fit, which can be assessed for both. Only then can we assess if the more complex autoregressive model offers any significant improvement over the simple linear regression. One such measure of fit is the Log-likelihood. It expresses the joint probability of the observed data given the model parameter, with the highest value (least negative) indicating the better model. Unlike the R-squared, Log-likelihood values cannot be easily interpreted alone because they are a function of sample size. However, they are well suited to compare the fit of different coefficients or models based on the same dataset.

We found the multiple linear regression model to be significant in the case of gastronomy, shopping, and work activity type (model p-value in Table 30). The empirical data suggest that pedestrian movement is not a significant predictor of variation in administrative, educational, and health activity intensity. From the two explanatory variables, the endogenous movement was a significant predictor of all three activity types, while the exogenous movement was the only significant predictor of work activity. From the three significant models, the pedestrian movement was able to explain 63% of the variance in gastronomy, 58% of the variance in shopping, and 40% of the variance in work activity intensity (adjusted R^2 in Table 30).

Table 30. Linear regression model performance and significance level.

	Administrative	Education	Gastronomy	Health	Shopping	Work
R^2	0.04	0.06	0.66	0.08	0.61	0.41
Adjusted R^2	-0.17	0.06	0.63	0.08	0.58	0.40
Log-likelihood	-25.91	-638.53	-45.25	-552.85	-60.58	-283.63
P-value (model)	0.12	0.17	$8.03e^{-7}$***	0.09	$6.32e^{-7}$***	$2.2e^{-16}$***
p-value exo. movement	0.93	0.07	0.99	0.12	0.08	2e-16***
p-value endo. movement	0.51	0.21	0.00***	0.08	0.00***	$1.1e^{-10}$***

Based on the significance of the regression models, for all further evaluations and testing, we consider only the effect of pedestrian movement on gastronomy, shopping, and work activity intensity.

Estimated Model Coefficients and Omitted Variable Bias

We visualize the coefficients for each variable by constructing their respective regression lines while holding the other variable constant at its mean. To emphasize the effect of omitted variable bias, we show the slope of the regression line for each variable as it would be fitted on its own. By doing so, we visualize the difference between coefficients estimated by the biased model and the true coefficients. These lines make not only clearer what was already visible from the fitted regression coefficients - namely, the direction in which endogenous and exogenous movement influence the activity levels, they also make clear the major impact of ignoring one of them.

The somewhat paradoxical effect of adding new variables on the coefficients of the old ones is also known as the Simpson paradox (Simpson, 1951). It is just another demonstration of the omitted variable bias occurring when we ignore a variable that is correlated with the explanatory part of the model. The aim of this study is to test for the occurrence of omitted variable bias when explaining the distribution of activities only by exogenous movement component (i.e., urban form). We expect that including the endogenous movement into the model does not only increase its precision (i.e.., R-squared, or Log-likelihood); more importantly, it significantly changes the estimated effect of exogenous movement on the allocation of activities. If this is the case, it means that ignoring the endogenous movement results in a biased model.

As discussed in Chapter 2.1, all models are wrong. So why should we care in this case as the apparent advantage of the simpler model is that it requires only one explanatory variable which can be directly derived from urban form? We argue that every model explains some portion of the reality and leaves the rest unexplained. Consequently, it is always wrong as it inherently contains an error - the unexplained portion. However, there is a difference between the model being wrong just because it is a simplification of reality and the model being wrong because the explained portion of reality is misleading. For practical reasons, we call the correct but incomplete model a *True* model and the incomplete but false model a *Biased* model. From an epistemological perspective, absolute knowledge is not feasible, and thus not knowing everything is not a problem but a necessity. However, it is a problem if something that we consider as knowledge turned out to be wrong. For this reason, we test if the omitted variable bias by comparing the coefficients for the model considering both movement components and the model considering only the exogenous movement.

$$M1(restricted) = \alpha_0 + \alpha_1 movement_{exogenous} + \mu_1 \quad (58)$$

$$M2(simultaneous) = \alpha_0 + \alpha_1 movement_{exogenous} + \alpha_2 movement_{endogenous} + \mu_2 \quad (59)$$

In addition to comparing the coefficients, we compare the model performance by running the ANOVA Chi-squared test for nested regression model testing the null hypothesis of their equality (i.e., there is no difference in the residual sum of squares). Consequently, if p-value < 0.05, we reject the null hypothesis and consider the restricted *Model 1* being significantly worse in explaining the variance in activity intensity than the simultaneous *Model 2*.

For all three remaining activities, the ANOVA Chi-squared test was highly significant (see Table 31). This means that removing the endogenous movement from the regression models causes a significant drop in the ability of the model to explain the variance in activity intensity.

When comparing the regression coefficient between the restricted and simultaneous model, we found a significant difference for all three activity types[47]. Consequently, we confirm the omitted variable bias caused by ignoring the endogenous movement when estimating the effect of pedestrian movement on gastronomy, shopping, and work activity intensity.

Table 31. ANOVA Chi-squared test for the nested regression model. A significant test and change in the estimated coefficients confirm the presence of omitted variable bias.

	Gastronomy	Shopping	Work
Model 1 - Individual estimation Exogenous movement coefficient (std. error)	1.86 (0.38)	1.94 (0.42)	0.57 (0.06)
Model 2 - Simultaneous estimation Exogenous movement coefficient (std. error)	-0.00 (0.57)	-1.75 (0.98)	0.06 (0.08)
ANOVA p-value	$1.17e^{-4}$ ***	$6.29e^{-5}$ ***	$1.10e^{-11}$ ***

We illustrate the impact of the omitted variable bias on the estimated regression coefficients by plotting the regression lines as fitted by the restricted and the simultaneous model (Figure 135). It is clearly visible that the estimated effect of exogenous movement on the allocation of activities does not only change in its size but also in the direction. As a result, such a model is inherently bias and not reliable.

[47] We calculate the 95% confidence interval (i.e. coefficient +- 2std.error) for the coefficients estimated by restricted and simultaneous model and look if they intersect. If they do not intersect, we consider them as significantly different.

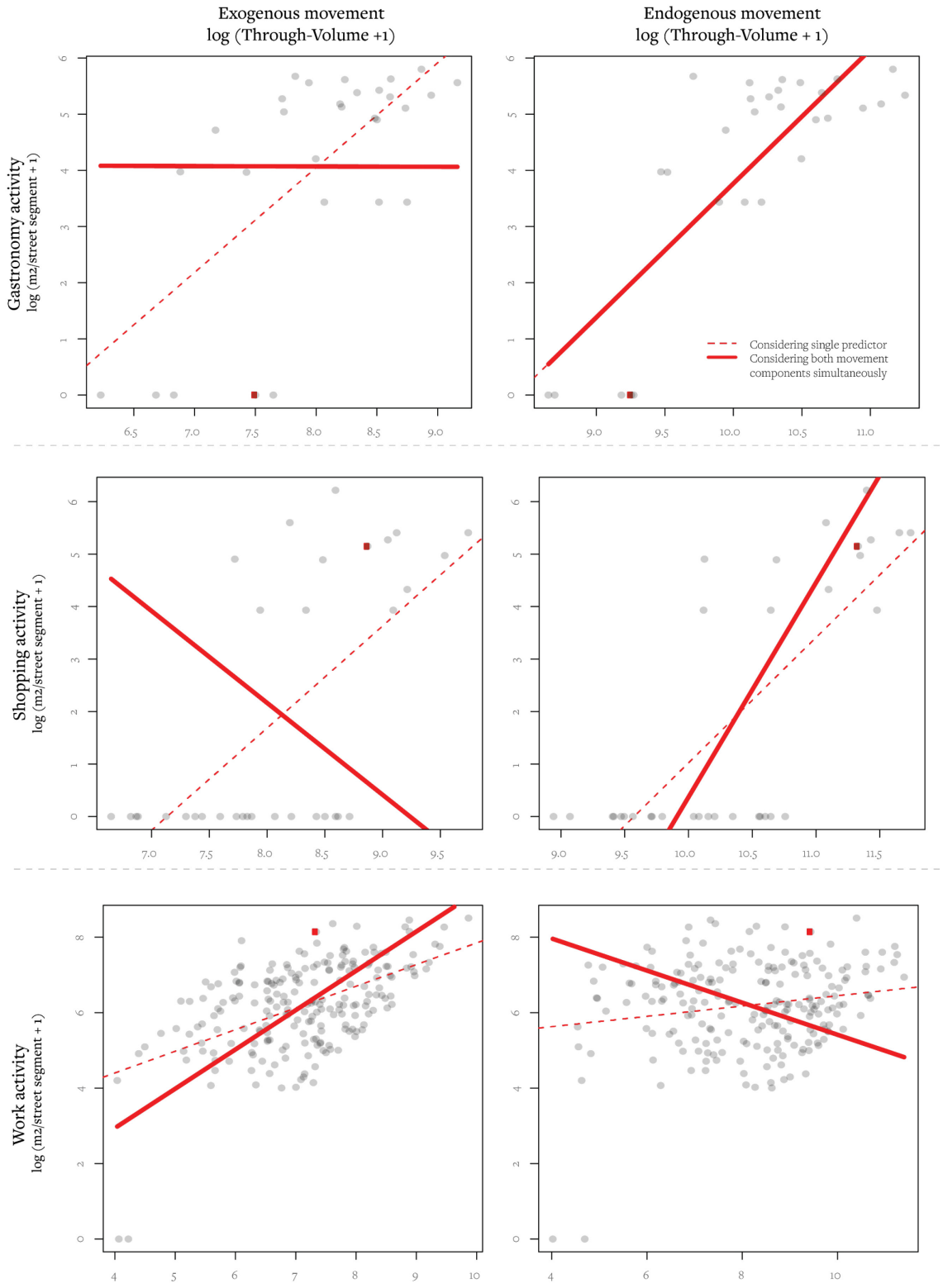


Figure 135. Regression line capturing the relationship between exogenous and endogenous movement and activity intensity when estimated simultaneously and individually. The regression lines for each variable are constructed by holding the other variable constant at its mean. The difference in the slope represent the amount of the omitted variable bias introduced by the individual estimation.

Non-spatial Diagnostics

Residual Plot

A large number of regression model diagnostics utilize residuals as a way of detecting potential misspecification issues. Residuals are particularly important as they represent what the model was not able to explain. Formally residual is the observed (y) – predicted ($\hat{y}$) values and can be both positive and negative.

$$residual = y - \hat{y} \tag{60}$$

In linear regression, this unexplained part of the model is assumed to be random and normally distributed. Thus, finding patterns in residuals means that the regression model might be flawed and should be reformulated. We test the residual normality via the Shapiro-Wilk test for normal distribution. The null hypothesis of the test assumes normality. Accordingly, if the test's p-value is below the critical level (p-value < 0.05) we reject the null hypothesis and conclude that the residual distribution is significantly different from the normal distribution. We visualize the regression residuals by a) plotting the predicted against the measured activity levels and by b) Quantile-Quantile plot showing deviations from the normal distribution. In the former, a correctly specified model yields a random pattern of both negative (red lines) and positive (blue lines) residuals. In the latter, we compare the residual quantiles with the theoretical quantiles of the normal distribution. If the residuals perfectly follow the normal distribution, then all points on the plot lie on the 45-degree red line.

After running the Shapiro-Wilk test, we found the residual distribution not significantly different from the normal distribution in the case of gastronomy and shopping activities (Table 32). The test result has been confirmed by the respective Quantile-Quantile plots (Figure 136b), with most data points lying close to the 45-degree line.

In the case of the work activity, the test statistic of the Shapiro-Wilk test for normality is $\chi^2 = 0.92$ with resulting highly significant p-value = $8.49e^{-9}$. Since this is below the critical threshold of 0.05, we reject the null hypothesis of normality. Consequently, we prefer the alternative and consider the residuals as being non-normally distributed. When investigating the residual plots (Figure 136a), we conclude that the only strong outliers from the normal distribution are the observations with measured zero-level activities. These outliers are associated with positive (i.e., non-zero) movement volume, which in turn results in the expectation of non-zero activity levels. This limited capability of linear regression to deal with the zero activity levels is known and was addressed by the two-step mixed model. We first filtered the zero-level activates by logistic regression model and then use the reduced dataset to fit the linear regression. Nevertheless, since not all zero-level observations could be filtered out, the residuals of the linear model tend to deviate from the normal distribution.

We must note that the non-normality does not cause any bias in the coefficient estimation, and as long as we are able to diagnose the source of this irregularity, the model is considered valid.

Table 32. Shapiro-Wilk test of normality of the multiple linear regression model residuals.

	Gastronomy		Shopping		Work	
	statistics	p-value	statistics	p-value	statistics	p-value
Shapiro-Wilk test	0.96	0.53	0.97	0.55	0.92	$8.49e^{-9}$ ***

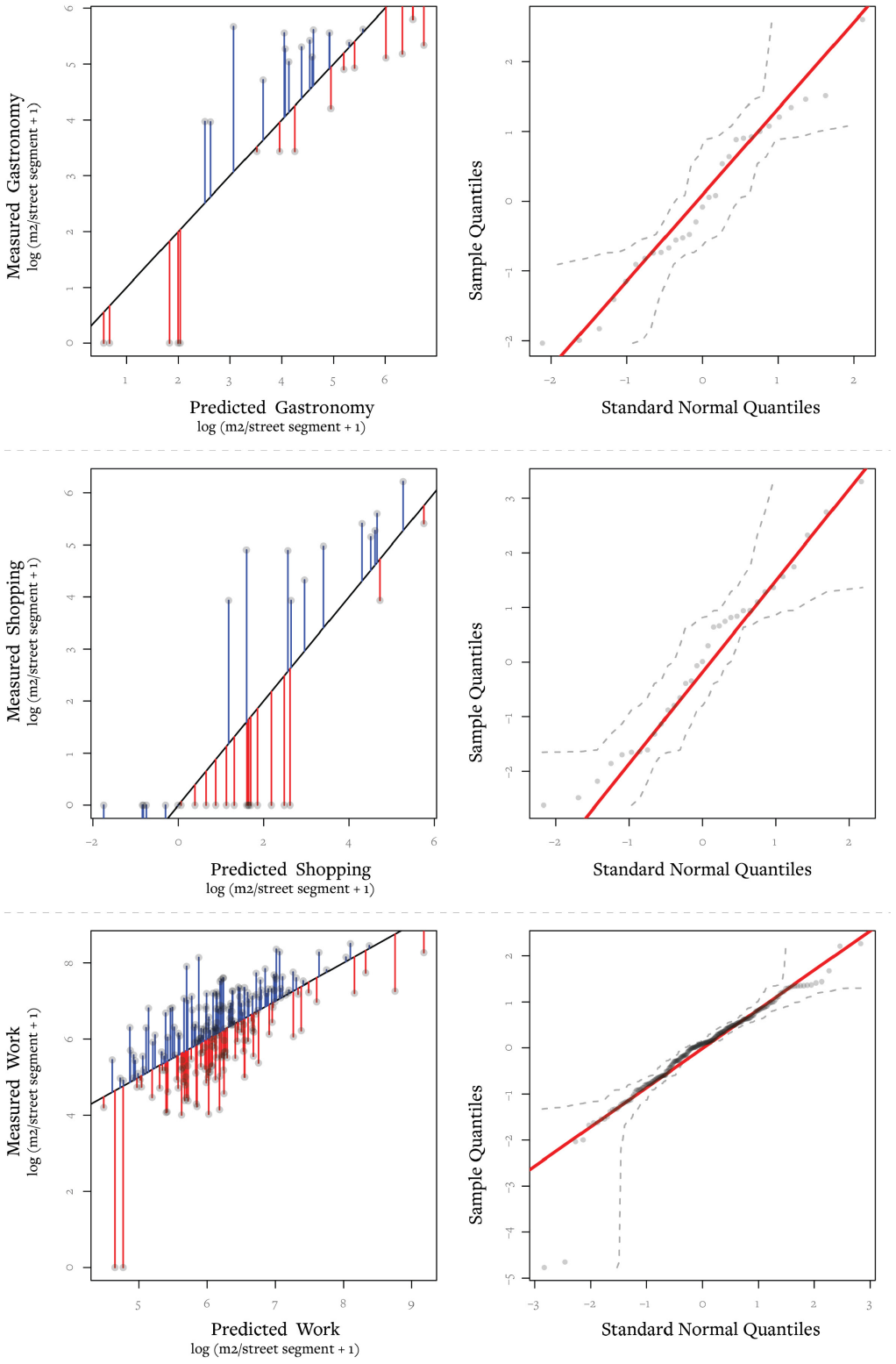


Figure 136. Residuals diagnostics plots. We visualize the measured vs. predicted activity intensity and Quantile-Quantile plot for gastronomy, shopping, and work activity. (red color = negative residuals, blue color = positive residuals).

Collinearity & Variance Inflation Factor

Collinearity, or in other words, correlated explanatory variables, is a common cause of instability in estimated model coefficients. As a result, it is not possible to reliably interpret the model in terms of understanding the individual effect of each explanatory variable. Colinear models might still be acceptable if their only purpose is prediction. However, in this study, the emphasis lies on interpreting the coefficients as we want to know if both exogenous and endogenous movement patterns are required to explain the activity levels and what role they play. The common measure of collinearity and indicator of potential model instability is the Variance Inflation Factor (VIF). Various recommendations for acceptable levels of VIF have been published in the literature. Perhaps most commonly, a value of 10 has been recommended as the maximum level of VIF (Hair et al., 1995; Neter et al., 1989). However, a recommended maximum VIF value of 5 (Rogerson, 2019) and even 4 (Pan & Jackson, 2008) can be found in the literature.

We found that the VIF for all three models was below the critical threshold of 10 (see Table 33). Consequently, we consider the regression models for gastronomy, shopping, and work as stable and reliable.

Table 33. Variation Inflation Factor in the multiple linear regression model caused by collinearity in the explanatory variables - exogenous and endogenous movement components.

	Gastronomy	Shopping	Work
Variance Inflation Factor	3.42	8.14	2.45

It might come as a surprise that the same explanatory variables produce models with different levels of collinearity. However, it is important to realize that for each activity type activity, we filter out different observations in the first step (i.e., *Filter* model). Therefore, we are looking at the same two variables but three different data sets. We demonstrate this by plotting the relationship of exogenous and endogenous movement for the street segments selected by the *Filter* model (Figure 137). We see that each plot differs in the strength of the linear relationship (R^2) and the number of observations.

We note that the strong correlation between the exogenous and endogenous movement component support the previous results finding the presence of omitted variable bias when one of the movement components is ignored in the process of estimating the allocation of activities. By definition, the regression model is biased when variables are excluded from the model are correlated with one or more explanatory variables included in the model.

We conclude that the presence of collinearity in the explanatory variables is strong enough to cause omitted variable bias, but it is l under the critical threshold, causing instability of the simultaneous model considering both variables.

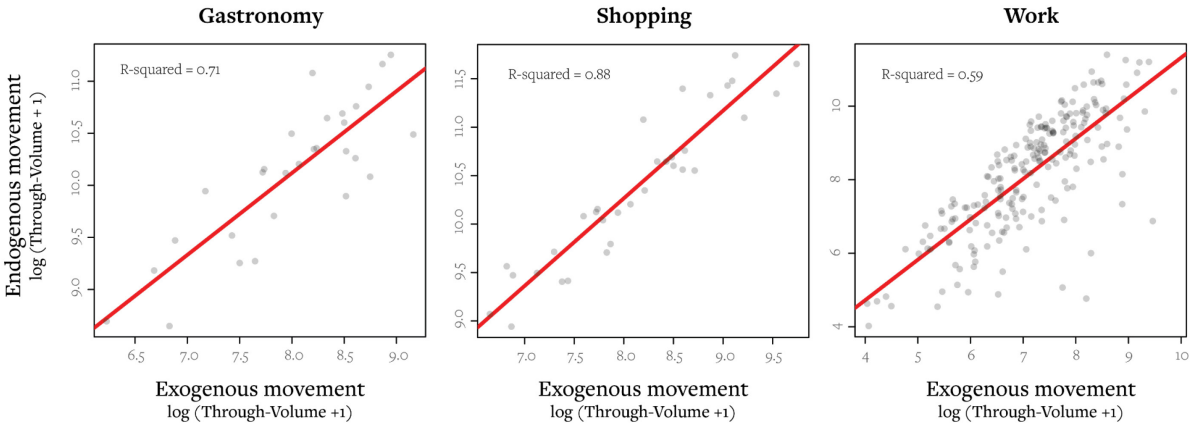


Figure 137. Collinearity between exogenous and endogenous movement components. We plot the observations identified in the Filter model as none-zero a) gastronomy activity street segments, b) shopping activity street segments, and c) work activity street segments.

Spatial Diagnostics

To test if there is a spatial dependence in the dependent variable (i.e., activity intensity) and thus, if the linear regression should control for this dependence, we run a set of spatial diagnostic tests. Common diagnostics to assess the presence of spatial dependence consist of two types of tests: diffused and focused tests. The main difference is that the diffuse test, also called a non-constructive test, is able to detect that there is a problem in the specification of the regression model but does not tell us what kind of problem it is. On the other hand, the focused test is testing specific types of possible problems and hence reveal not only the source but also the solution to a specification problem.

Based on this logic, we run both types of tests in a sequence. First, we look at the diffuse test to reject the absence of spatial autocorrelation, and then if necessary, we use the focused test to detect the source of the autocorrelation. We note that spatial autocorrelation can appear due to various reasons and not only as a product of spatial interaction - spillover effect, which is the primary hypothesis tested in this study. Spatial autocorrelation can also be a product of spatial heterogeneity or spatial error autocorrelation (Anselin & Rey, 1991). Without going into details, the point is that different processes can yield similar outcomes and thus need specific tests.

Arguably, the most commonly applied diffused test statistics for spatial autocorrelation is Moran's I. It was first applied to the regression residual by Cliff and Ord (Cliff & Ord, 1972), and due to its power against many alternatives, it is also known as misspecification test. After running the diffuse test and rejecting the null hypothesis of no spatial autocorrelation, we move to the focused tests. These tests are designed to reject the null hypothesis against a fully specified alternative model, such as the spatial lag model (i.e., the spatial dependence comes from spillover between activities) or spatial error model (i.e., the spatial dependence come from somewhere else, we don't know from where but if we model

it we gain on precision). The focused tests always compare the constrained to the unconstrained model, and if they are significantly different, we reject the null hypothesis. The constrained model means simply that the spatial term ρ is constrained to 0 (i.e., in case of spacial lag, this would result in simple linear regression), and the unconstrained model is a fully specified model controlling for spatial dependence. After running the constrained and unconstrained model, there are three classic approaches to compare them. We can compare the value of the model estimates (Wald test), their fit (Likelihood ratio test), or the slope of the likelihood function (Lagrange Multiplier test). All these options are asymptotically equivalent, which means that for large samples, they yield the same results. However, for smaller samples, the Wald test is the least, and the Lagrange multiplier is the most conservative test in terms of rejecting the null hypothesis (i.e., $WA \geq LR \geq LM$). Moreover, the Lagrange Multiplier test is simpler to estimate, and thus, for the following model specification and selection procedure, we adopt the most conservative out of the tree tests - the Lagrange multiplier and follow the specification decision rules laid out by Ansellin and Rey (2014).

Constrained model when testing for spatial lag:

$$y = \rho W y + \alpha X + \mu \tag{61}$$

$$H_0: \rho = 0 \tag{62}$$

The Global Moran I statistics for regression residuals reveals significant autocorrelation in the case of gastronomy and work activities (Table 34). We found diffusion (i.e., negative autocorrelation) in gastronomy and clustering (i.e., positive autocorrelation) in work activities. In the case of shopping activity, we failed to reject the null hypothesis, which means that no significant autocorrelation was detected.

Table 34. Unfocused test for spatial autocorrelation of unknown type.

	Gastronomy		Shopping		Work	
	statistics	p-value	statistics	p-value	statistics	p-value
Moran's I residual autocorrelation test	-0.29	0.04 *	0.17	0.09	0.14	0.01 **

The Lagrange Multiplier focused tests for spatial lag and spatial error suggests the spatial lag alternative for both autocorrelated activity types – gastronomy and work (Table 35). In the case of shopping activity, the focused test confirms the results of unfocused Moran's I test with no significant autocorrelation being detected.

Table 35. Focused test for spatial autocorrelation of spatial lag and spatial error type.

	Gastronomy		Shopping		Work	
	Chi-square (df = 1)	p-value	Chi-square (df = 1)	p-value	Chi-square (df = 1)	p-value
Lagrange Multiplier Spatial Lag	4.23	0.03 *	$5e^{-3}$	0.93	2.08	0.14
Lagrange Multiplier Spatial Error	2.18	0.13	0.90	0.34	4.12	0.045 *

Unconstrained Model - Spatial Auto-regressive Model

Based on the results of the spatial diagnostics of the linear regression model, we extend the LM model in the case of gastronomy and work activities by the autoregressive term to account for the autocorrelation in activities.

Diagnostics

Estimated Model Coefficients

From the two movement components, we found only the endogenous movement being a highly significant predictor of gastronomy activity. Furthermore, we conclude that the spatial autoregressive coefficient was significant and negative with rho = -0.38 and p-value = 0.03. This implies negative spatial autocorrelation, or in other words, an increase in Gastronomy activities at a given location is related to a decrease in Gastronomy activities at neighboring locations. It must be noted that at first glance, the dispersion of gastronomy activities might look counterintuitive; we argue that it is only a matter of the data aggregation unit. As we discuss in the limitation section, the unit of aggregation might significantly influence, even reverse the trend due to the phenomena known as the Modifiable Area Unit Problem (MAUP). In essence, we might see different patterns and trends based on the analysis scale. It might be possible that there is a clustering of gastronomy activities on a building or block scale and, at the same time, dispersion on the street or neighborhood scale.

In the case of the work activity, we found both, exogenous and endogenous movement components being highly significant. Like the LM model, each movement component coefficient is showing in a different direction. The increase in the exogenous movement was associated with an increase in work activity intensity while everything else being equal. On the other hand, the increase in the endogenous movement was associated with a decrease in Work activity intensity. We conclude that the spatial auto-regressive coefficient was

significant and positive, with the spatial auto-regressive coefficient Rho = 0.07 (p-value = 0.045). This implies positive spatial autocorrelation, or in other words increase in work activity intensity at a given location triggers an increase in Work activity intensity at neighboring locations.

Table 36. Spatial auto-regressive model performance and significance levels.

	Gastronomy	Work
Log-likelihood	-43.02	-280.65
P-value (model)	$2e^{-3}$ ***	$1e^{-10}$ ***
p-value exo. movement	0.2	2e-16 ***
p-value endo. movement	$2e^{-7}$ ***	$3e^{-10}$ ***
p-value spatial auto-regressive term	0.03 *	0.04 *

Log-likelihood Ratio

The Log-likelihood ratio Chi-squared test is a hypothesis test used to choose the best model between two nested models. By nested models, we mean that one model is a special or constrained case of the other. In our case, the spatial lag SARLM model is a special case of a simple linear regression LM model. The null hypothesis is that both models are equally good. If the null hypothesis is rejected ($p < 0.05$), then the extended model is a significant improvement over the simpler model.

As matter of fact, this is what we described earlier as a focused Likelihood ratio test of spatial dependence. By large samples, it should gain the same results as the Lagrange Multiplier test. However, in our case, the two Log-likelihoods are computed from models that use different estimators. The linear regression is estimated via OSL, and SARLM is estimated via the Maximum Likelihood. In the case of the spatial dependence likelihood ratio test, both models are estimated with the same estimator.

We found that the Log-likelihood ratio test confirmed the Lagrange Multiplier test and proved the SARLM model as being a significant improvement over the LM for both gastronomy and work activity (Table 37).

Table 37. Log-likelihood ratio test

	Gastronomy		Work	
	Chi-square (df = 1)	p-value	Chi-square (df = 1)	p-value
Log-likelihood ratio test	4.45	0.03 *	5.97	0.01 **

Spatial Multiplier, Direct and Indirect Effects

In a simple linear regression model, the effect of change in explanatory variable x on dependent variable y is constant, and it equals the coefficient α. However, this is not true in the spatially dependent case as captured in Equation (63. Here, due to the spatial dependence and resulting spillovers, the total effect of change in x is more than in non-spatial settings. This total effect is also called the spatial multiplier. It consists of the direct effect and the indirect effect. The direct effect corresponds to the effect found in simple linear regression, while the indirect effect is generated by the interaction in the neighborhood structure (see Equation (64 and (65). By separating these two effects, we quantify how changes in the explanatory variables at one location indirectly influence its neighborhood.

$$total\ effect = [y|\triangle X] = (I - \rho W)^{-1}(\triangle X)\alpha \quad (63)$$

$$direct\ effect = (\triangle X)\alpha \quad (64)$$

$$indirect\ effect = [(I - \rho W)^{-1} - I](\triangle X)\alpha \quad (65)$$

We found a strong indirect effect of spatial autocorrelation, causing almost half of the total effect on gastronomy activity intensity (Table 38. Direct and indirect effects.). In the case of the work activity intensity, the effect of spatial autocorrelation was still significant but at a much lower level of 7% of the total effect.

Table 38. Direct and indirect effects.

	Gastronomy	Work
Direct effect	50.61%	92.81%
Indirect effect	49.39%	7.19%

Appendix 19 Combined Activity Prediction Model (Filter + Amplifier)

In the following, we present the results of the combined *Filter* and *Amplifier* model of the pedestrian movement effect and autocorrelation on the allocation of activities. From the six activity types considered in this study, we found the gastronomy, shopping, and work activities being significantly affected by pedestrian movement. We run the combined *Filter* and *Amplifier* model to estimate the effect of exogenous and endogenous movement on each of the three significant activity types. Finally, we compare the estimated activity intensity with the observed one to assess the model prediction accuracy (Figure 138).

Figure 138. Spatial distribution of prediction residuals by activity type (blue = positive residuals, red = negative residuals).

Residuals

The residual standard error is arguably one of the most useful characteristics of the model goodness of fit (Equation 66). It provides the absolute measure of the typical distance that the data points fall from the regression line. It tells us how precise our prediction is directly in the units of the dependent variable. Assuming normal distribution of the residuals, the standard error can also be converted into a 95% prediction interval (PI). By calculating the standard error and its prediction intervals, we identify the upper and lower bounds of our prediction error. In our case, we calculate the 95% confidence interval for maximum and minimum error in the predicted activity floor area per street segment. We argue that this measure is not only possible to calculate for any regression model but is also easier to interpret and offers a practical way how to assess if a model is good enough.

A good indicator of the model performance is also a comparison of the residual standard error to the standard deviation of the data. If these two are similar, it means that the prediction model performs only as good as guessing the mean.

$$Standard\ error = \sqrt{\frac{\sum r^2}{n - (1 + k)}} \tag{66}$$

With n as the number of observations, r as the regression residuals and k as the number of predictors.

To account for the different capacities of each street to accommodate activities, we also calculate the relative prediction error. This is simply the average deviation between predicted and observed activities in percentage.

$$Relative\ mean\ error = \frac{\sum |r|}{n} \tag{67}$$

We found the predicted shopping and gastronomy to be on average no more than 0.25% and 1.18% different from the actual value, while the prediction of work activities is with the mean error of 14% percent less accurate (Table 39. Performance of the combined model (i.e., Filter + Amplifier).). The high prediction accuracy can be attributed to the fact that most streets have zero-activity intensity, and the ability of the *Filter* to correctly identify them dramatically increases the performance of the combined activity prediction model.

Table 39. Performance of the combined model (i.e., Filter + Amplifier).

	Gastronomy	Shopping	Work
Standard deviation	66 m^2	41 m^2	641 m^2
Standard error	42 m^2	15 m^2	373 m^2
Relative mean error	1.18%	0.25%	14.20%

In Figure 141, Figure 144, and Figure 147, we explore the spatial distribution of the prediction residual by activity type. In general, residuals represent factors affecting the activity intensity, which were not considered in the prediction model. In all three activities, we find the positive residuals (i.e., observed activity intensity is lower than the predicted) concentrated around the historic city center. On the contrary, negative residuals are mostly scattered all over the study area and tend to occupy the city center. We note that the goal of this study is not to explain the additional factors causing that potential for activity intensity is not fully realized or exceed our expectations. Nevertheless, we argue that the activity-movement model resented here makes it possible to systematically investigate such phenomena and its causes.

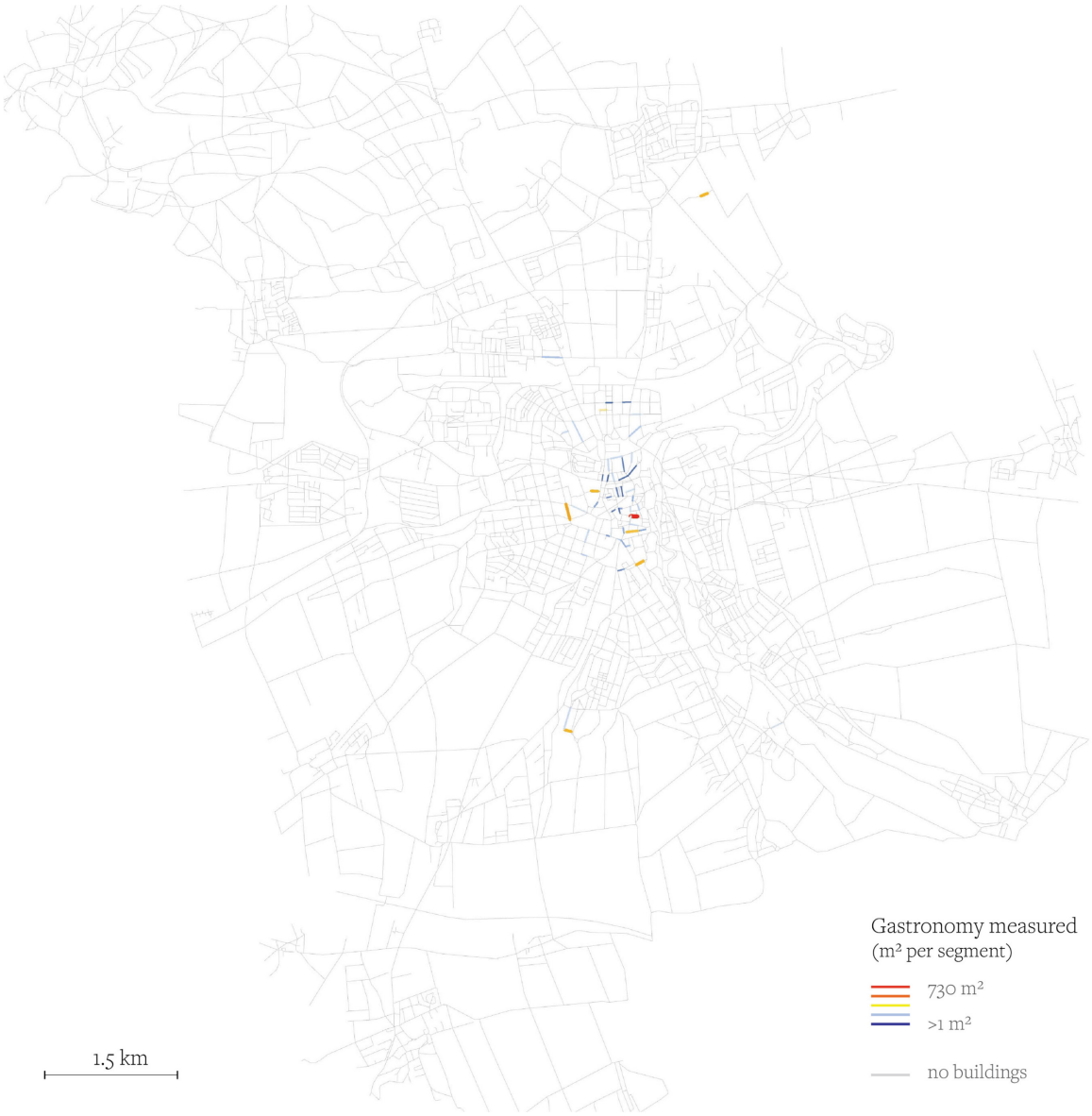


Figure 139. Measured Gastronomy activity intensity per street segment.

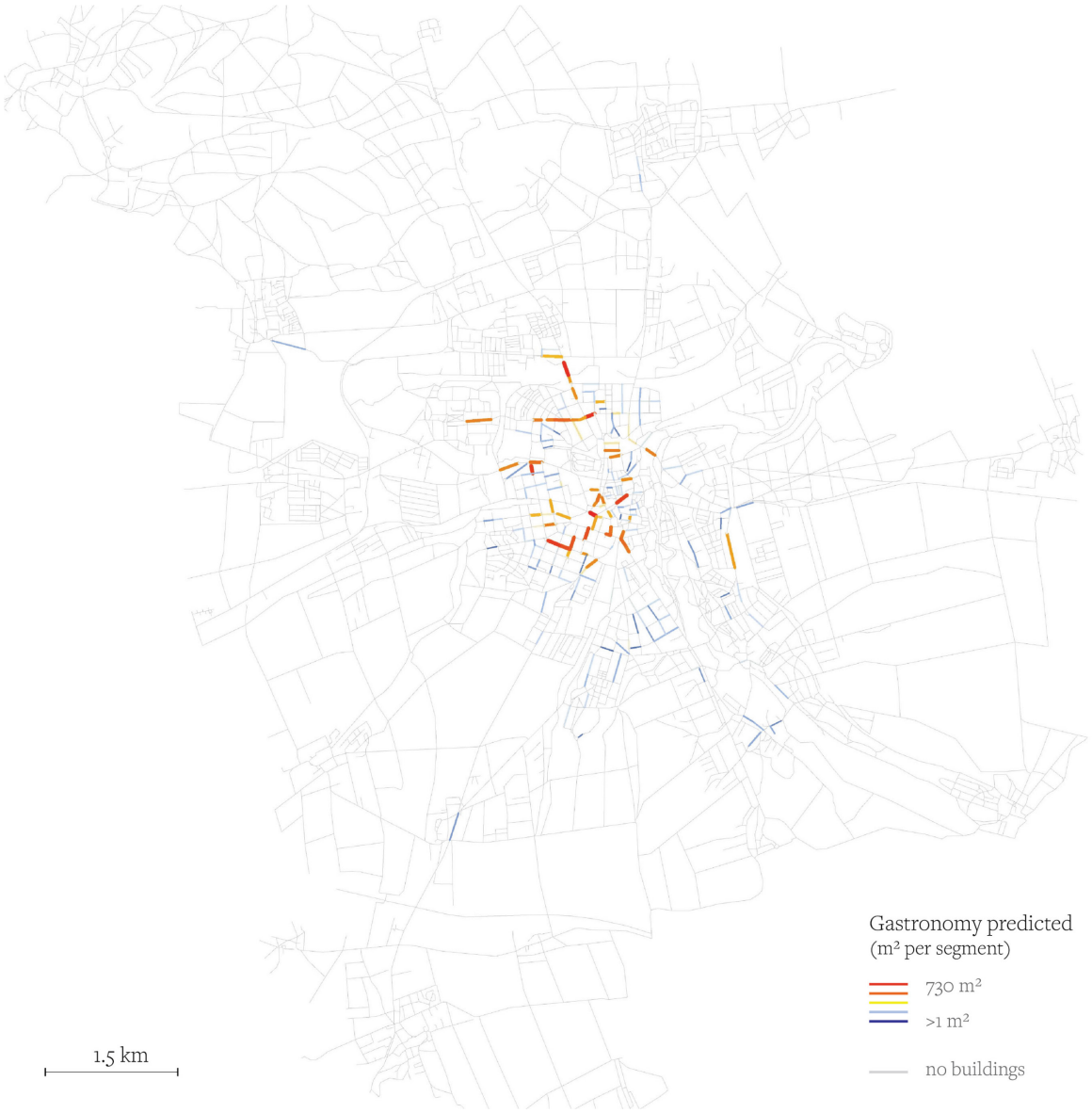


Figure 140. Predicted Gastronomy activity intensity per street segment (combined Filter and Amplifier model).

Figure 141. Gastronomy activity prediction residuals (combined Filter and Amplifier model).

Figure 142. Measured Shopping activity intensity per street segment.

Figure 143. Predicted Shopping activity intensity per street segment (combined Filter and Amplifier model).

Figure 144. Shopping activity prediction residuals (combined Filter and Amplifier model).

Figure 145. Measured Work activity intensity per street segment.

Figure 146. Predicted Work activity intensity per street segment (combined Filter and Amplifier model).

Figure 147. Work activity prediction residuals (combined Filter and Amplifier model).

Summary

We preset the comprehensive overview of the activity prediction model attributes. We list both sub-models (i.e., *Filter* and *Amplifier*) as well as the combined model (Table 40. Summary of Filter, Amplifier, and Mixed model of the pedestrian movement effect and spatial autocorrelation on activity allocation.). We conclude that pedestrian movement is a significant predictor of a street with zero-level activities of all types. Based on the activity type, we were able to correctly identify and filter out 27% to 83% of zero-level activity streets.

When it comes to the prediction of the activity intensity at a street that passes the *Filter*, the pedestrian movement was significant only in the case of gastronomy, shopping, and work activities. Additionally, we conclude that the estimation of the activity intensity by using only the exogenous movement leads to omitted variable bias. Finally, we tested for spatial dependence in activities and detect positive autocorrelation in work and negative autocorrelation in gastronomy activities.

Table 40. Summary of Filter, Amplifier, and Mixed model of the pedestrian movement effect and spatial autocorrelation on activity allocation.

			Administrative	Education	Gastronomy	Health	Shopping	Work
Filter	AUROC		0.73	0.65	0.81	0.84	0.91	0.75
	Sensitivity		94%	90%	90%	92%	90	90%
	Specificity		27%	30%	64%	55%	83%	34%
	Misclassification error		71%	67%	34%	44%	17%	36%
	Exo. Movement		/	+	-	-	-	+
	Endo. Movement		+	/	+	+	+	-
Amplifier	R^2		/	/	0.66	/	0.58	0.41
	Exo		/	/	/	/	/	+
	Endo		/	/	+	/	+	-
	Omitted variable bias		/	/	YES	/	YES	YES
	Residual Normality		/	/	YES	/	YES	NO
	Collinearity (VIF>10)		/	/	NO	/	NO	NO
	Spatial dependence	Lag	/	/	YES	/	NO	YES
		Error	/	/	NO	/	NO	NO
	Indirect effect		/	/	-49.39%	/	/	7.19%
Combined model	Relative mean error (%)		/	/	1.18%	/	0.25%	14.20%

References

Alfonzo, M. A. (2005). To walk or not to walk? The hierarchy of walking needs. *Environment and Behavior*, 37(6), 808–836.

Alonso, W. (1964). *Location and Land Use: Toward a General Theory of Land Rent*. Harvard University Press.

Amemiya, T. (1985). *Advanced econometrics*. Cambridge, Mass. : Harvard University Press. http://archive.org/details/advancedeconomet00amem

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Springer Science & Business Media.

Anselin, L. (2001). Spatial econometrics. *A Companion to Theoretical Econometrics*, *310330*.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, *89*(1), 3–25.

Anselin, L. (2013). *Spatial econometrics: Methods and models* (Vol. 4). Springer Science & Business Media.

Anselin, L, & Rey, S. (1991). Properties of tests for spatial dependence in linear regression models. *Geographical Analysis*, *23*(2), 112–131.

Anselin, L, & Rey, S. (2014). *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC.

Barceló, J. (2010). Models, Traffic Models, Simulation, and Traffic Simulation. In J. Barceló (Ed.), *Fundamentals of Traffic Simulation* (pp. 1–62). Springer.

Batty, M. (2009). *Notes on Accessibility In Search of a Unified Theory Invited Paper 2*. 1–4.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, *22*(6), 725–730.

Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral Science*, *10*(2), 161–163.

Ben-Akiva, M., & Lerman, Y. (1979). Disaggregate travel and mobility choice models and measures of accessibility. In *Behavioural travel modelling* (pp. 654–679). Croom Helm.

Berghauser Pont, M. Y., & Haupt, P. A. (2007). The relation between urban form and density. *Urban Morphology*, *11*(1), 62–65.

Berry, B. J. L. (1967). *Geography of Market Centers and Retail Distribution*. Prentice-Hall.

Bhat, C., Handy, S., Kockelman, K. M., Mahmassani, H., Weston, L., Gopal, A., & Srour, I. (2002). Development of an Urban Accessibility Index: A Summary. *Project Summary Report*, *4938-S*, 4.

Bhat, C., Handy, S., Kockelman, K., Mahmassani, H., (2002). Development of an Urban Accessibility Index: Formulations, Aggregation, and Application. *Time*, *7*(21), 176.

Bielik, M., Emo, B., Schneider, S., & Hölscher, C. (2017). Does urban density follow centrality? Empirical study on the influence of street network centrality on urban density and its

implications for the prediction of pedestrian flows. *Proceedings - 11th International Space Syntax Symposium, SSS 2017*. 11th Space Syntax Symposium, Lisbon, Portugal.

Bielik, M., König, R., Fuchkina, E., Schneider, S., & Abdulmalik, A. (2019). *Evolving Configurational Properties—Simulating multiplier effects between land use and movement patterns*. 12th Space Syntax Symposium, Beijing, China.

Blake, P. (1977). *Form follows fiasco: Why modern architecture hasn't worked*. Little, Brown Boston.

Bloch, P. H., Ridgway, N. M., & Nelson, J. E. (1991). Leisure and the shopping mall. *ACR North American Advances*.

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). Academic Press.

Brueckner, J. K. (2003). Strategic interaction among governments: An overview of empirical studies. *International Regional Science Review*, *26*(2), 175–188. Scopus.

Cascetta, E., Cartenì, A., & Montanino, M. (2013). A New Measure of Accessibility based on Perceived Opportunities. *Procedia - Social and Behavioral Sciences*, *87*, 117–132. https://doi.org/10.1016/j.sbspro.2013.10.598

Cascetta, E., Cartenì, A., & Montanino, M. (2016). A behavioral model of accessibility based on the number of available opportunities. *Journal of Transport Geography*, *51*, 45–58.

Cervero, R., & Center, R. (1988). *America's Suburban Centers: A Study of the Land Use-Transportation Link*. United States. Dept. of Transportation. Technology Sharing Program.

Chiaradia, A., Hillier, B., Barnes, Y., & Schwander, C. (2009). *Residential property value patterns in London: Space syntax spatial analysis*.

Christaller, W. (1933). *Die zentralen Orte in Süddeutschland: Eine ökonomisch-geographische Untersuchung über die Gesotzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen*. Gustav Fischer.

Clarke, K. A. (2005). The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science*, *22*(4), 341–352.

Cliff, A., & Ord, K. (1972). Testing for Spatial Autocorrelation Among Regression Residuals. *Geographical Analysis*, *4*(3), 267–284.

Cormen, T. H., Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction To Algorithms*. MIT Press.

Cournot, A. (1838). Recherches sur les principes mathematiques de la theorie des richesses (Hachette, Paris).

Cullen, I., & Godson, V. (1975). Urban networks: The structure of activity patterns. *Progress in Planning*, *4*, 1–96.

Cunningham, H. G. (2004). *The Simplest Thing that Could Possibly Work*.

Curtis, C., & Scheurer, J. (2010). Planning for sustainable accessibility: Developing tools to aid discussion and decision-making. *Progress in Planning*, *74*(2), 53–106.

Cuthbert, A. R. (2007). Urban design: Requiem for an era–review and critique of the last 50 years. *Urban Design International*, *12*(4), 177–223.

Dalton, R. C. , & Dalton, N. (2007). *Applying Depth Decay Functions to Space Syntax Network Graphs*.

Dalton, N. (2001). Fractional Configurational Analysis and a solution to the Manhattan problem. *3rd International Space Syntax Symposium*, *January 2001*, 26:1-13.

Dalton, R. C. (2003). The secret is to follow your nose: Route path selection and angularity. *Environment and Behavior*, *35*(1), 107–131.

De Jong, G., Daly, A., Pieters, M., & van der Hoorn, T. (2007). The logsum as an evaluation measure: Review of the literature and new results. *Transportation Research Part A: Policy and Practice*, *41*(9), 874

Detzer, D. (2002). *Allegiance: Fort Sumter, Charleston, and the Beginning of the Civil War*. Harcourt.

DiPasquale, D., & Wheaton, W. C. (1996). *Urban economics and real estate markets* (Vol. 23). Prentice Hall Englewood Cliffs, NJ.

Doherty, S. T. (2006). Should we abandon activity type analysis? Redefining activities by their salient attributes. *Transportation*, *33*(6), 517–536.

Dudey, M. (1993). A note on consumer search, firm location choice, and welfare. *The Journal of Industrial Economics*, 323–331.

Eaton, B. C., & Lipsey, R. G. (1975). The Principle of Minimum Differentiation Reconsidered: Some New Developments in the Theory of Spatial Competition. *The Review of Economic Studies*, *42*(1), 27–49.

Elldér, E. (2014). Residential location and daily travel distances: The influence of trip purpose. *Journal of Transport Geography*, *34*, 121–130.

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. Cambridge University Press.

Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design*, *14*(1), 65–84.

Frank, L. D., Sallis, J. F., Saelens, B. E., Leary, L., Cain, K., Conway, T. L., & Hess, P. M. (2010). The development of a walkability index: Application to the Neighborhood Quality of Life Study. *British Journal of Sports Medicine*, *44*(13), 924–933.

Frankenstein, J. (2015). *Human spatial representations and spatial memory retrieval* [Doctoral Thesis, ETH Zurich].

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, *40*(1), 35–41.

Fujita, M., Krugman, P. R., & Venables, A. (1999). *The spatial economy: Cities, regions and international trade*. MIT Press.

Fujita, M., Thisse, & Thisse, J.-F. (2002). *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*. Cambridge University Press.

Fox, M. (1995). Transport planning and the human activity approach. Journal of Transport Geography, 3(2), 105–116.

Gale, N., Golledge, R. G., Pellegrino, J. W., & Doherty, S. (1990). The acquisition and integration of route knowledge in an unfamiliar neighborhood. *Journal of Environmental Psychology*, *10*(1), 3–25.

Garrison, W. L. (1960). Connectivity of the Interstate Highway System. *Papers in Regional Science*, *6*(1), 121–137.

Gauthier, P., & Gilliland, J. (2005). *Mapping urban morphology: A classification scheme for interpreting contributions to the study of urban form*. 10.

Gehl, J. (1987). Life Between Buildings: Using Public Space. In *The City Reader*. van Nostrand Reinhold.

Gehl, J. (2014). *Istanbul – Public Space Public Life*.

Gibbons, S., & Overman, H. G. (2012). Mostly Pointless Spatial Econometrics?*. *Journal of Regional Science*, *52*(2), 172–191.

Gil, J. (2015). *Examining ' edge e ffects ': Sensitivity of spatial network centrality analysis to boundary conditions*. 186–189.

Golledge, R. G. (1995). Path Selection and Route Preference in Human Navigation. *Transportation*, *277*.

Gorter, C., & Nijkamp, P. (2001). Location Theory. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 9013–9019). Pergamon.

Grasser, G., Van Dyck, D., Titze, S., & Stronegger, W. (2013). Objectively measured walkability and active transport and weight-related outcomes in adults: A systematic review. *International Journal of*

Greene, W. H. (2017). *Econometric Analysis*. Pearson Education.

Gröger, G., & Plümer, L. (2012). CityGML – Interoperable semantic 3D city models. *ISPRS Journal of Photogrammetry and Remote Sensing*, *71*, 12–33.

Hägerstrand, T. (1970). What about people in Regional Science? *Papers of the Regional Science Association*, *24*(1), 6–21.

Hair, J., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). Multivariate data with readings. *US America: Prentice Hall Inc.*

Handy, S., & Clifton, K. (2001). Evaluating Neighborhood Accessibility: Possibilities and Practicalities. *Journal of Transportation and Statistics*, 67–78.

Handy, S., & Niemeier, D. A. (1997). Measuring accessibility: An exploration of issues and alternatives. *Environment and Planning A*, *29*(7), 1175–1194.

Handy, S. (1992). *Regional versus Local Accessibility: Variations in Suburban Form and the Effects on Non-Work Travel*; University of California Transportation Center, Working Papers). University of California Transportation Center.

Handy, S. (2002). *Accessibility- vs. Mobility-Enhancing Strategies for Addressing Automobile Dependence in the U.S.*

Handy, S. (2009). Accessibility- vs. Mobility-Enhancing Strategies for Addressing Automobile Dependence in the U.S. *Institute of Transportation Studies*, *Issues in*, 57–66.

Hansen, W. G. (1959). How Accessibility Shapes Land Use. *Journal of the American Institute of Planners*, *25*(2), 73–76.

Hanson, S. (1995). Getting there: Urban transportation in context. *Geography of Urban Transportation*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media.

Helbing, D. (1998). From microscopic to macroscopic traffic models. In J. Parisi, S. C. Müller, & W. Zimmermann (Eds.), *A Perspective Look at Nonlinear Media* (pp. 122–139). Springer.

Hillier, B. (1996). Cities as movement economies. *Urban Design International*, 1(1), 41–60.

Hillier, B. (1998). *Space is the Machine: A Configurational Theory of Architecture*. Cambridge University Press.

Hillier, B. (1999). The hidden geometry of deformed grids: Or, why space syntax works, when it looks as though it shouldn't. *Environment and Planning B: Planning and Design*, *26*(2), 169–191.

Hillier, B., Burdett, R., & Penn, A. (1987). *Creating Life: Or, Does Architecture Determine Anything ? 250*, 233–250.

Hillier, B., & Hanson, J. (1984). *The Social Logic of Space*. Cambridge University Press.

Hillier, B., & Iida, S. (2005). Network and psychological effects in urban movement. *Proceedings of the 5th International Symposium on Space Syntax*, *1*(1987), 475–490.

Hillier, B., Penn, A., Grajewski, T., & Xu, J. (1993). Natural Movement—Or, Configuration and Attraction in Urban Pedestrian Movement. *Environment and Planning B: Planning and Design*, *20*(1), 29–66.

Hise, R. T., Kelly, J. P., Gable, M., & McDonald, J. B. (1983). Factors Affecting the Performance of Individual Chain Store Units-An Empirical-Analysis. *Journal of Retailing*, *59*(2), 22–39.

Hochmair, H., & Frank, A. (2002). Influence of estimation errors on wayfinding-decisions in unknown street networks–analyzing the least-angle strategy. In *Spatial Cognition and Computation* (Vol. 2, Issue Hochmair 2000).

Holton, R. H. (1958). The Distinction between Convenience Goods, Shopping Goods, and Specialty Goods. *Journal of Marketing*, *23*(1), 53–56.

Ingene, C. A. (1984). Structural determinants of market potential. *Journal of Retailing*, *60*(1), 37–64.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. Random House.

Jiang, B. (2015). Geospatial analysis requires a different way of thinking: The problem of spatial heterogeneity. *GeoJournal*, *80*(1), 1–13.

Jones, P. M., Dix, M. C., Clarke, M. I., & Heggie, I. G. (1983). Understanding travel behaviour.

Kansky, K. J. (1963). *Structure of transportation networks: Relationships between network geometry and regional characteristics.* University of Chicago, Department of Geography.

Kant, E. (1933). Ümbrus, majandus ja rahvastik Eestis. *Ökoloogilismajandusgeograafiline Uurimus "(Doktoritöö, Tartu Ülikool).*

Kant, Edgar. (1935). *Bevölkerung und Lebensraum Estlands: Ein anthropoökologischer Beitrag zur Kunde Baltoskandias*. Akadeemiline kooperatiiv.

King, J., & Murphy, M. (2017). Walking—The Silver Bullet to Health and Wellbeing (breakout presentation). *Journal of Transport & Health*, *7*, S44.

Kolbe, T. H. (2009). Representing and Exchanging 3D City Models with CityGML. In J. Lee & S. Zlatanova (Eds.), *3D Geo-Information Sciences* (pp. 15–31). Springer Berlin Heidelberg.

Krenz, K. (2017). *Employing Volunteered Geographic Information in Space Syntax Analysis*. 11th International Space Syntax Symposium, Lisbon, Portugal

Krier, L. (2009). *The Architecture of Community*. Island Press.

Kropf, K. (1996). Urban tissue and the character of towns. *Urban Design International*, *1*(3), 247–263.

Kwan, M.-P. (1998). Space-Time and Integral Measures of Individual Accessibility: A Comparative Analysis Using a Point-based Framework. *Geographical Analysis*, *30*(3), 191–216.

Lee, J. H., & Ostwald, M. J. (2019). *Grammatical and Syntactical Approaches in Architecture: Emerging Research and Opportunities*. IGI Global.

Lefebvre-Ropars, G., Morency, C., Singleton, P. A., & Clifton, K. J. (2017). Spatial transferability assessment of a composite walkability index: The Pedestrian Index of the Environment (PIE). *Transportation Research Part D: Transport and Environment*, *57*(October), 378–391.

Lehmann, E. L. (1951). A General Concept of Unbiasedness. *Annals of Mathematical Statistics*, *22*(4), 587–592.

Lerman, Y., Rofé, Y., & Omer, I. (2014). Using space syntax to model pedestrian movement in urban transportation planning. *Geographical Analysis*, *46*(4), 392–410.

Lerner, A. P., & Singer, H. W. (1937). Some Notes on Duopoly and Spatial Competition. *Journal of Political Economy*, *45*(2), 145–186.

Li, Y., & Tsukaguchi, H. (2005). Relationships Between Network Topology and Pedestrian Route Choice Behavior. *The Eastern Asia Society for Transportation Studies*, *6*(1), 241–248.

Marshall, A. (1925). *Principles of Economics*. Ravenio Books.

Marshall, S. (2005). *Streets & patterns* (1st ed). Spon.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396.

Miller, H. (2007). Place-based versus people-based geographic information science. *Geography Compass*, *1*(3), 503–535.

Minsky, M. (1968). Matter, Mind and Models. *Semantic Information Processing*, *1*.

Mishra, S. K. (2007). A brief history of production functions. *Available at SSRN 1020577*.

Mitchell, G. (1993). The practice of operational research. *Systems Research*, *12*(2), 179–179.

Morris, J. M., Dumble, P. L., & Wigan, M. R. (1979). Accessibility indicators for tansport planning. *Transportation Research Part A: Policy and Practice*, *13*(2), 91–109.

Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models*.

Netto, V. M. (2016). 'What is space syntax not?'Reflections on space syntax as sociospatial theory. *Urban Design International*, *21*(1), 25–40.

Nevin, J. R., & Houston, M. J. (1980). Image as a component of attraction to intraurban shopping areas. *Journal of Retailing*, *56*(1), 77–93.

Nourian, P., Hoeven, F. V. D., Rezvani, S., & Sariylidiz, S. (2015). Easiest paths for walking and cycling: Combining syntactic and geographic analyses in studying walking and cycling mobility. *Proceedings of the 10th International Space Syntax Symposium*, 1–15.

Nourian, P., Rezvani, S., Valeckaite, K., & Sariyildiz, S. (2018). Modelling walking and cycling accessibility and mobility: The effect of network configuration and occupancy on spatial dynamics of active mobility. *Smart and Sustainable Built Environment*, *7*(1), 101–116.

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122.

Oliveira, V. (2016). *Urban morphology: An introduction to the study of the physical form of cities*. Springer.

Pafka, E., Dovey, K., & Aschwanden, G. D. (2020). Limits of space syntax for urban design: Axiality, scale and sinuosity. *Environment and Planning B: Urban Analytics and City Science*, *47*(3), 508–522.

Pan, Y., & Jackson, R. T. (2008). Ethnic difference in the relationship between acute inflammation and serum ferritin in US adult males. *Epidemiology and Infection*, *136*(3), 421–431.

Peeta, S., & Ziliaskopoulos, A. K. (2001). Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Networks and Spatial Economics*, *1*(3), 233–265.

Penn, A., & Dalton, R. C. (1994). The architecture of society: Stochastic simulations of urban pedestrian movement. In *Simulating Societies* (Vol. 5). UCL press.

Peponis, J., Dalton, N., & Dalton, R. (2003). *To Tame a TIGER one has to know its nature*. 65.

Piketty, T. (2017). *Capital in the Twenty-First Century*. Harvard University Press.

Pinho, P., & Oliveira, V. (2009). *Different approaches in the study of urban form: Journal of Urbanism: International Research on Placemaking and Urban Sustainability: Vol 2, No 2*.

Plackett, R. L. (1949). A Historical Note on the Method of Least Squares. *Biometrika*, *36*(3/4), 458–460. JSTOR.

Popper, K. R. (1959). The Logic of Scientific Discovery. University Press.

Porta, S. (2010). Network in Urban Design. Six Years of Research in Multiple Centrality Assessment. In *Network Science* (Issue October 2015, pp. 107–129). Springer London.

Porta, S., Strano, E., Iacoviello, V., Messora, R., Latora, V., Cardillo, A., Wang, F., & Scellato, S. (2009). Street Centrality and Densities of Retail and Services in Bologna, Italy. *Environment and Planning B: Planning and Design*, *36*(3), 450–465.

Primerano, F., Taylor, M. A. P., Pitaksringkarn, L., & Tisato, P. (2008). Defining and understanding trip chaining behaviour. *Transportation*, *35*(1), 55–72.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Raford, N., Chiaradia, A., & Gil, J. (2007). *Space Syntax: The Role of Urban Form in Cyclist Route Choice in Central London*.

Ratti, C. (2004). Space syntax: Some inconsistencies. *Environment and Planning B: Planning and Design*, *31*(4), 487–499.

Renn, A. M. (2010, August 1). *The Mark of a Great City Is in How It Treats Its Ordinary Spaces, Not Its Special Ones*.

Renn, A. M. (2013). *The Urban State of Mind: Meditations on the City*.

Reyer, M., Fina, S., Siedentop, S., & Schlicht, W. (2014). Walkability is only part of the story: Walking for transportation in Stuttgart, Germany. *International Journal of Environmental Research and Public*

Roberson, R. (2016). Enlightened Piety during the Age of Benevolence: The Christian Knowledge Movement in the British Atlantic World. *Church History*, *85*(2), 246–274.

Rodríguez, D. A., & Joo, J. (2004). The relationship between non-motorized mode choice and the local physical environment. *Transportation Research Part D: Transport and Environment*, *9*(2), 151–173.

Rogerson, P. A. (2019). *Statistical Methods for Geography: A Student's Guide*. SAGE.

Ruskin, J. (1853). *The Stones of Venice*. Dent.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.

Scheer, B. C. (2015). Epistemology of Urban Morphology. *Urban Morphology*.

Sevtsuk, A. (2010). *Path and place: A study of urban geometry and retail activity in Cambridge and Somerville, MA* [Dissertation, Massachusetts Institute of Technology].

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238–241.

Smith, A. (1776). An Inquiry Into the Nature and Causes of the Wealth of Nations. Whitestone.

Sober, E. (1988). Likelihood and Convergence. *Philosophy of Science*, *55*(2), 228–237.

Soja, E. (2001). In different spaces. *Proceedings of the 3rd International Symposium on Space Syntax Georgia Institute of Technology, Atlanta, GA, Pp S1*, 1–s1.

Speck, J. (2012). *Walkable City: How Downtown Can Save America, One Step at a Time*. Farrar, Straus and Giroux.

Takeuchi, D. (1977). A study on pedestrian route choice behavior. *Proceedings of JSCE,* 295.

Taneja, S. (1999). Technology Moves In. *Chain Store Age*, *75*(5), 136–137.

Thünen, J. H. (1826). *Der isolirte Staat in Beziehung auf Landwirtschaft und Nationalökonomie, oder, Untersuchungen über den Einfluss den die Getriedepreise, der Reichthum des Bodens und die Abgaben auf den Ackerbau ausüben*. Wirtschaft & Finan.

Tobler, W. R. (1970). A Computer Movie Simulation Urban Growth in Detroit Region. *Economic Geography*, *46*, 234–240.

Troped, P., Saunders, R., Pate, R., Reininger, B., Ureda, J., & Thompson, S. (2001). Associations between Self-Reported and Objective Physical Environmental Factors and Use of a Community Rail-Trail. *Preventive Medicine*, *32*(2), 191–200.

Turner, A., & Dalton, N. (2005). A simplified route choice model using the shortest angular path assumption. *Geocomputationorg*.

Turner, A. (2001). Angular analysis. Proceedings of the 3rd International Symposium on Space Syntax, 30–1.

Turner, A. (2007). From Axial to Road-Centre Lines: A New Representation for Space Syntax and a New Model of Route Choice for Transport Network Analysis. *Environment and Planning B: Planning and Design*, *34*(3), 539–555.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

Tversky, B. (1993). Cognitive Maps, Cognitive Collages, and Spatial Mental Models in. In Andrew U. Frank & I. Campari (Eds.), *SPATIAL INFORMATION THEORY A THEORETICAL BASIS FOR GIS* (Vol. 716, Issue SEPTEMBER 1993, pp. 1–17). Springer-Verlag.

Vale, D. S., & Pereira, M. (2016). The influence of the impedance function on gravity-based active accessibility measures: A comparative analysis. *Environemnt and Planning B*.

Vale, D. S., Saraiva, M., & Pereira, M. (2015). Active accessibility: A review of operational measures of walking and cycling accessibility. *Journal of Transport and Land Use*, *1*(JANUARY), 1–27.

Varoudis, T., Law, S., Karimi, K., Hillier, B., & Penn, A. (2013). Space syntax angular betweenness centrality revisited. *2013 International Space Syntax Symposium*.

Vernengo, M., Caldentey, E. P., & Rosser Jr, B. J. (Eds.). (2020). *The New Palgrave Dictionary of Economics*. Palgrave Macmillan UK.

Weber, A. (1909). *Über den Standort der Industrien*.

Whyte, W. H. (1988). *City: Rediscovering the Center*. Doubleday.

Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils, by Philip G. Wright, with the Aid of the Council and Staff of the Institute of Economics*. Macmillan Company.

Zhang, W.-B. (2002). *An Economic Theory of Cities: Spatial Models with Capital, Knowledge, and Structures*. Springer-Verlag Berlin Heidelberg.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley Press.

List of Figures

List of Tables

Abbreviations

A	Activity
ANOVA	Analysis of Variance
AUROC	Area Under the Receiver Operating Characteristic
BLUE	Best Linear Unbiased Estimator
CI	Confidence Interval
CUM	Configurational Urban Morphology
DTA	Dynamic Traffic Assignment
en	Endogenous
ex	Exogenous
FA	Factor Analysis
FAMI	Form Activity Movement Interaction
FPR	False Positive Rate
KS	Kolmogorov–Smirnov Test
LM	Linear Regression Model
LM	Lagrange Multiplier Test
LoD	Level of Detail
LR	Likelihood Ratio Test
M	Movement
MiD2017	Mobilität in Deutschland Study 2017
OD	Origin-Destination
OGC	Open Geospatial Consortium
OLS	Ordinary Least Squares
OSM	OpenStreetMap
PCA	Principal Component Analysis
PSP2017	Pedestrian Shortest Path Study 2017
SARLM	Spatial Autoregressive Linear Model
SPP	Spatial Point Process
TP	Transportation lanning
TPR	True Positive Rate
U	Urban Form
UE	Urban Economy
VIF	Variance Inflation Factor
WA	Wald Test
WAI	Walkability Index
WS	Walk Score

Mathematical Notation

The mathematical notation used throughout this dissertation follows the vector notation if not explicitly defined otherwise. This means that variables representing vectors (one dimensional array of numbers) are noted as capital letter of Latin or Roman alphabet. Coefficients are noted in lowercase Greek alphabet.

List of symbols:

A	Total activity intensity
A_T	Activity intensity for activity type T
$A_{T(en,ex)}$	Endogenous or exogenous component of activity intensity for activity type T
$A_{T(en,ex)sim}$	Simulated endogenous or exogenous component of activity intensity for activity type T
M	Total movement
M_T	Movement to the activity type T
$M_{T(en,ex)}$	Endogenous or exogenous component of movement to the activity intensity for activity type T
$M_{T(en,ex)sim}$	Simulated endogenous or exogenous component of movement to the activity intensity for activity type T
I	Identity matrix
μ	Error term
ρ	Spatial auto-regressive coefficient
$\hat{Y}$	Hat symbol represent predicted value or variable
R^2	R-squared
χ^2	Chi-squared
W	Spatial weights matrix
$\sum$	Summation

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle unmissverständlich gekennzeichnet.

Weitere Personen waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder anderen Personen) in Anspruch genommen. Niemand hat von mir unmittelbar der mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Ich versichere ehrenwörtlich, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Ort, Datum

Unterschrift