# Climate Change and Geographic Information in Real Estate Research

A DISSERTATION SUBMITTED TO THE GRADUATE FACULTY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR RERUM POLITICARUM (DR. RER. POL.)

SUBMITTED TO

THE FACULTY OF BUSINESS, ECONOMICS AND MANAGEMENT INFORMATION SYSTEMS OF

THE UNIVERSITY OF REGENSBURG

SUBMITTED BY

Jens Hirsch

DIPLOM-GEOGRAPH, UNIVERSITY OF REGENSBURG

ADVISORS:

PROF. DR. SVEN BIENERT

PROF. DR. GREGOR DORFLEITNER

**Table of content**


1 Introduction ..... 4
1.1 General Motivation ..... 4
1.2 Research Questions ..... 8
1.3 Course of Analysis ..... 9
1.4 References ..... 10
2 Energy efficiency: Behavioural effects of occupants and the role of refurbishment for European office buildings ..... 12
2.1 Introduction ..... 13
2.2 Background and empirical framework ..... 14
2.3 Dataset ..... 17
2.4 Econometric approach ..... 22
2.5 Research results ..... 24
2.6 Conclusion ..... 29
2.7 References ..... 32
3 Assessment of climatic risks for real estate ..... 35
3.1 Introduction ..... 36
3.2 Background ..... 37
3.3 Conceptualization of the ImmoRisk tool ..... 38
3.4 Technical implementation of the ImmoRisk tool ..... 44
3.5 Functional implementation of the ImmoRisk tool ..... 48
3.6 Risk assessment results of the ImmoRisk pilot study ..... 52
3.7 Conclusion ..... 54
3.8 References ..... 55
3.9 Appendix ..... 57
4 The analysis of customer density, tenant placement and coupling inside a shopping centre ..... 59
4.1 Introduction ..... 60
4.2 Literature review: shopping centre tenant mix ..... 60
4.3 GIS-techniques ..... 62
4.4 Conclusions and discussion ..... 81
4.5 References ..... 84
5 Variable Clumping Method and Mean-k-Nearest-Neighbor Method- Introducing two new approaches to retail concentration measurement to shopping center research ..... 87
5.1 Introduction ..... 88
5.2 Literature Review: Concentration of Retail Categories ..... 88
5.3 Variable Clumping method ..... 89
5.4 Mean-k-Nearest-Neighbor ..... 98
5.5 Conclusion ..... 99
5.6 References ..... 101
6 Do urban tourism hotspots affect Berlin housing rents? ..... 102
6.1 Introduction ..... 103
6.2 Literature review ..... 104
6.3 Data ..... 105
6.4 Methodology ..... 106
6.5 Empirical results ..... 112

6.6 Conclusion ..... 121
6.7 References ..... 122
6.8 Appendix ..... 126
7 Conclusion ..... 129
7.1 Executive Summary ..... 129
7.2 Final Remarks ..... 133
7.3 References ..... 134

# 1 Introduction

## 1.1 General Motivation

### 1.1.1 Climate Change an its economic effects

Climate change is one of the main challenges presently facing the real estate industry. There is an urgent need for action, because climate change is already taking place and will accelerate in the coming decades. July 2015 was the warmest month in the existing weather records dating back to 1880[1] and the atmosphere is heating up faster than over the last 1,000 years[2]. The first of the above facts refutes the thesis of climate change 'taking a break' during the last decade. The second illustrates that the observed warming in the 20th and 21st century is not a 'natural' phenomenon of increasing temperatures after the end of the last glacial period, as often claimed by 'climate skeptics'. Instead, there is strong consensus amongst climate scientists that global warming is anthropogenic. A recent study by Cook et al. (2013, p. 3) found that 98.4% of climate scientists endorsed this consensus.

The so-called Stern Review, authored by the former World Bank Chief Economist Sir Nicolas Stern for the British government, was the first comprehensive study on the global economic impacts of climate change. The results were rather pessimistic, with an estimated risk of losing up to 20% of global *GDP* each year, if climate change and its consequences are not tackled early enough.[3] The good news was that a relatively moderate share of only one percent of global *GDP* per year would need to be invested in order to avoid the worst impacts of climate change. However, a recent study from Burke et al. (2015) states that the consequences might be even worse than predicted in the Stern Review and estimates a probability of at least 40% (and even more in some scenarios) for a decline of over 20% of global *GDP*. The study also rejects the common thesis that wealthy countries will be significantly more capable of coping with climate change.

### 1.1.2 Coping with climate change: Mitigation and adaptation

Usually, a differentiation is made between two main strategies for dealing with climate change, namely mitigation and adaptation. Measures aiming at the prevention of global warming by reducing emissions of greenhouse gases are denoted as mitigation. By contrast, the term adaptation denotes the conversion of existing structures, aiming at better conformation to a changing climate. The affected structures involve infrastructure and buildings, as well as, for example, farming practices and social structures of health care and risk management.

The real estate industry is thought to be responsible for approximately 40% of global energy consumption and 20-30% of global greenhouse gas (*GHG*) emissions[4]. This role as a main driver of climate change implies a great responsibility on the part of the real estate industry to take measures of mitigation. Corresponding initiatives can spring from the industry itself or in the form of legal regulations like the German *EnEV* (*Energy Saving Ordinance*). Regulatory mitigation frameworks are becoming progressively tighter in virtually all countries and are a major driver of increasing production costs. Therefore, one question of scientific interest is the extent to which these tightened frameworks have resulted in a reduced energy consumption of newly constructed or refurbished buildings. Profound knowledge about the interaction between building characteristics and occupant behavior with respect to energy demand is essential, in

[1] NOAA, 2015.
[2] Smith et al., 2015.
[3] Stern, 2006.
[4] Cf. World Economic Forum, 2016 (20% GHG-share) and UNEP, 2009 (30%).

order to assess the effectiveness of these measures and the associated ecological and economic sustainability. The first paper of this dissertation ('Energy efficiency: Behavioral effects of occupants and the role of refurbishment for European office buildings') deals with this question regarding a sample of European office buildings provided by the *Green Rating Alliance* (*GRA*). The paper examines, among other questions, whether younger buildings and those that have undergone refurbishment consume less energy. Unexpectedly, young buildings and refurbished ones seem to have significantly *higher* energy consumption than other objects. Possible explanations of this apparent failure of recent regulations to reduce energy consumption are discussed, including the statistical interaction of refurbishment and building age. The paper further addresses some behavioral aspects of energy consumption, because the most energy-efficient physical building qualities become worthless, if the people who use it fail to save energy in practice. One main finding in opposition to our initial expectations is that buildings used by several different tenants consume less energy than single-tenant buildings. The relevance of multi-dimensional sustainability ratings with further sustainability indicators besides energy consumption are analyzed through the relation of modeled water and actual energy consumption.

### 1.1.3 Adapting the real estate industry to a changing climate

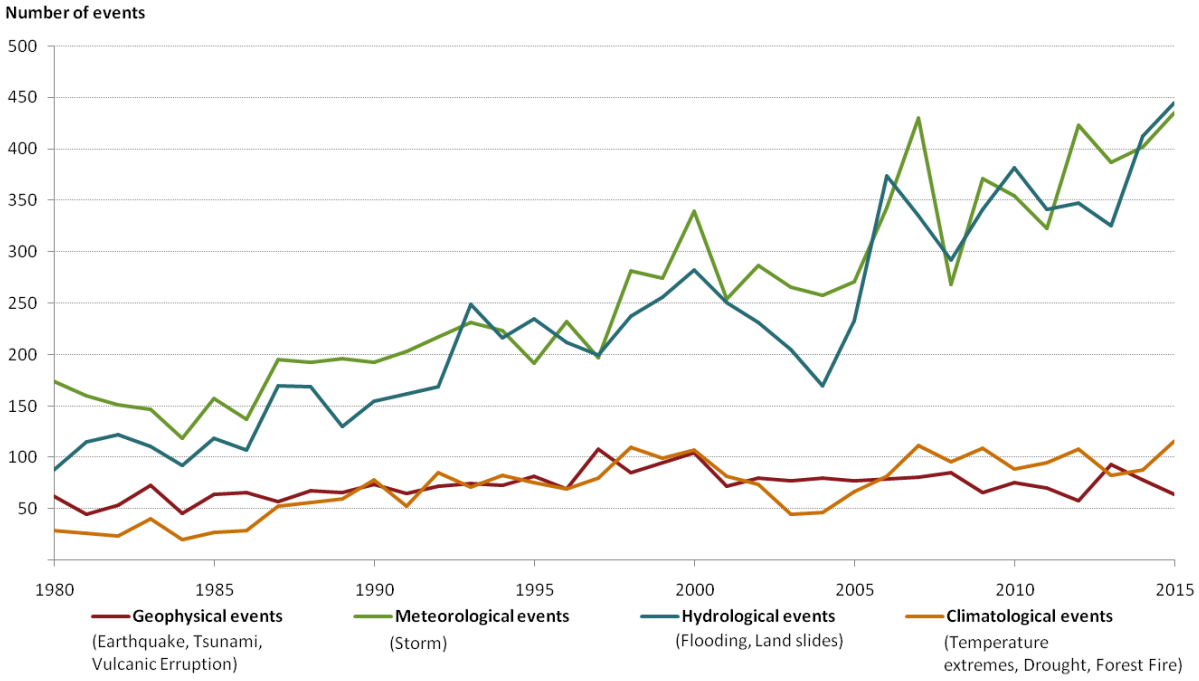
The real estate industry, however, is not only a major contributor to climate change, it is also seriously affected by its consequences. Grothmann et al. (2009) estimate that building damage will represent one quarter of climate-change-related losses and that approximately 2.5% of all investments in real estate will be necessary for an appropriate adaptation. Natural disasters cause damages to the building stock amounting to billions of Euros in Germany alone. The declining temperature difference between the Arctic and low latitudes will lead to an accumulation of weather conditions associated with heat waves, storms and flooding[5,6,7]. Increasing extreme weather events cause a wide range of direct, indirect and consequential losses[8] and require distinct measures of adaptation in terms of more resistant structures and improved risk management systems. As mentioned, climate change will accelerate over the course of the 21st century, but is already taking place right now. According to the world's largest reinsurance company *Munich Re*, the number of non-geophysical natural disasters has tripled since 1980 and annual mean damages have even quadrupled in the same period[9] (see Figure 1). As a consequence, the real estate industry is facing rising insurance premiums and in the worst case scenario, some regions might even completely lose their economic usability and viability, for example due to rising sea levels, frequent river flooding or extreme droughts.

Despite the obvious need for adaptation measures as soon as possible, the real estate industry still reacts mainly passively. The reason for this is a lack of ready-to-use information about the quantity of impacts and its spatial distribution. This uncertainty impedes an optimal allocation of capital to adaptation measures and endangers the economic sustainability of the entire industry. Climate data and damage models which are necessary for quantitative risk estimations are basically available, but access is complicated and, more particularly, the data are not combined in a way that would be integrable into existing real estate risk management systems. Accordingly, the second paper ('Assessment of climatic risks for real estate') describes the development of the *ImmoRisk*-web-tool, which is a first step towards overcoming the mentioned deficit of information. The tool addresses the material needs of relevant stakeholders like

[5] Coumou et al., 2014.
[6] Petoukhov et al., 2013.
[7] Screen/Simmonds, 2014.
[8] Bienert, 2014.
[9] Munich Re, GeoRisks Research, NatCatSERVICE. 2014.

homeowners, developers and communities, and overcomes many flaws in existing tools, by considering all relevant natural hazards in a quantitative manner, including present and future climate conditions, and delivering quantitative monetary risk measures at the object level. The tool is free of charge to users and integrates climate data about natural hazards with statistical damage models and an appraisal approach.

**Figure 1: Number and costs of natural disasters (worldwide 1980-2015)**




*Source: Munich Re (2014), Munich Re (2015), Munich Re (2016)*

The paper gives an overview of the final tool, including its programming by the authors, and describes the necessary steps of pre-processing climate data with extreme value statistical methods and how these hazard data are combined with damage functions and replacement costs into an annual expected loss. The evaluation of the risk calculations for the 15 covered locations indicate a significant spatial variation of expected risk measures and a particularly strong increase in storm damage due to climate change. The results demonstrate the importance of geographic analysis and its user-friendly presentation for the real estate industry.

### 1.1.4 Geographical analysis of customer behavior and tenant placement in shopping centers

With computers becoming more powerful and an increasing amount of available spatial data on various types of topics, new fields of analysis emerge and can help to improve all kinds of activities related to the real estate industry. The second paper illustrates the use of geographical data for an estimation of natural risks and their importance for efficiently coping with the challenges of climate change, which demand rapid adaptation. The third and the fourth paper ('The analysis of customer density tenant placement and coupling inside a shopping centre with GIS' and 'Variable Clumping Method and Mean-k-Nearest-Neighbor Method - Introducing two new approaches to retail concentration measurement to shopping center research') shift the focus onto structures and dynamics within shopping centers, still keeping an eye on the economic implications of spatial phenomena.

The management of shopping centers aims at maximizing rents and revenues. Besides the location of the shopping center and marketing activities, the spatial structure of the shopping center and its tenants are

the key success factors. Shopping centers are facing increasing competition with online shopping, and customers are nowadays much better informed, more sovereign and less loyal (Stepper, 2014). Therefore, center management requires detailed information about customer behavior in order to improve center performance. The search for an optimal tenant mix (number, nature and size of tenants, as well as the placement of their outlets) has produced a large number of studies dealing with potential success factors (Nelson, 1958; Brown, 1992; Carter and Haloupek, 2000; Yuo et al., 2004; Carter and Vandell, 2005; Des Rosiers et al., 2009). However, it remains difficult for the management of an individual center to incorporate the findings of this research. The same holds true for the analysis of coupling, which denotes a visit to two specific shops during one shopping trip. The phenomenon of coupling raises the question of whether shops from the same retail category should be concentrated or rather dispersed, in order to generate synergetic effects. GIS offers a wide range of possibilities for gaining the necessary information to help assess the status quo and implement consistent improvements. The use of advanced GIS techniques further enables the generation of more valid information that can be used as input for scientific regression studies on the factors influencing revenues and rents.

The two papers are based on a comprehensive study of a large German shopping center with about 55.000 m² sales area. The study comprises an on-site survey capturing the shopping behavior of over 1.000 customers, and an analysis of the tenant mix itself. Customers were interviewed directly at the exits after finishing their shopping trips, including questions about their walking paths within the center, shops visited and various socio-demographic data. By combining the gathered information and entering it into a GIS, it was possible to analyze coupling between different stores, visualize customer densities and make precise distinctions between different groups of visitors. These results enable the center management to identify problem areas with a low customer density, and potential improvements based on coupling behavior. The center management can further develop marketing activities adapted to specific target groups.

The fourth paper deals particularly with clustering of shops of the same retail category. The findings from different studies are not coherent concerning whether a dispersed (Carter and Haloupek, 2002) or a clustered arrangement (Yuo et al., 2004) is favorable. These inconsistent results may emerge partly from different definitions of when two or more shops should be actually regarded as clustered. We propose using of the *Variable Clumping Method,* which is based on a probabilistic approach and enables an analysis on different spatial scales. Its application enables statements to be made about the significance of certain clumps and has not yet been applied at all in the context of shopping centers or similar spatial arrangements. Furthermore, it could easily also be applied to the analysis of downtown retail agglomerations. Finally, the fourth paper introduces the *Mean-k-Nearest-Neighbor Method* to the shopping center research. It needs less computational power than the *Variable Clumping Method* and enables an easy-to-implement approach for an initial graphical identification of clusters from certain retail categories.

### 1.1.5 Effects of urban tourism hotspots on the Berlin rental market

The fifth paper ('Do urban tourism hotspots affect Berlin housing rents?') shifts the geographical focus away from the conditions within a certain building (the shopping center) onto the urban context. Here, the capabilities of GIS are used to generate geostatistical data for an analysis of the Berlin rental market focusing on the impact of urban tourism.

Overnight stays have increased by almost 10% per year since 2006 and reached approximately 27 million in 2014 (Eurostat, 2015). Although urban tourism plays a very important role in Berlin's economy, its consequences are controversial. There is a growing debate about the side-effects of increasing urban

tourism on the residential housing market in Berlin. Especially private short-term rentals as holiday flats are blamed for decreasing the housing supply illegitimately. The continuously growing demand for housing and an insufficient growth of supply is held liable for the increasing rents of recent years in Berlin. Some initial measures like rent control or the misuse-prohibition-law have already been taken to keep housing affordable.

The paper analyzes the role of urban tourism hotspots in the Berlin rental market, applying several *geoadditive* hedonic regression models drawing on market data from Germany's leading real estate portal *Immobilienscout24* and data on relevant amenities from the tourist information platform *Tripadvisor*. This unique dataset combines information about rental units, including rents, object features and location, with a comprehensive geodatabase of tourist attractions, hotels, restaurants and holiday flats, their location and importance. The main research question was whether the concentrations of tourist infrastructure ('urban tourism hotspots') affect housing rents of the surrounding rentals flats and whether different types of amenities differ with regard to their effects. Due to the large dataset, it was also possible to test the reasons for the recent rent increases, using the regression models. Here, the question is whether the increases can be explained, for example, in terms of better quality, or whether they are rather a market effect of growing demand.

## 1.2 Research Questions

The following sections summarize the research questions of the dissertation, subdivided into the five papers.

**Energy efficiency: Behavioural effects of occupants and the role of refurbishment for European office buildings**

- How do physical building characteristics and occupant attributes influence the energy consumption of European office buildings?
- Is there a significant difference between single- and multi-tenant buildings concerning energy consumption?
- How can *Generalized Additive Models* improve our understanding of drivers of energy consumption in office buildings?
- What is the effect of refurbishment measures and strengthened requirements for energy efficiency?
- Are there indications of a *rebound effect* in the analyzed sample of office buildings?

**Assessment of climatic risks for real estate**

- What is the current state of research with regard to the assessment of natural hazards and risks to real estate in Germany?
- What are the theoretical and practical requirements of a tool that can assess these natural risks and what might a final approach to its realization look like?
- Who are the main stakeholders that might from such a tool and what is its contribution to climate change adaptation?
- What relevant information about present and future natural hazards is available from regional climate models and how can this information be connected with property-specific information on vulnerability?
- How could climate and risk researchers further co-ordinate their activities?

- What is the effect of climate change on property-specific natural risks in Germany?

**The analysis of customer density, tenant placement and coupling inside a shopping centre with GIS**

- How can the application of *GIS* improve both theoretical research and the practical management of shopping centers?
- What are the advantages of a personal on-site customer survey for the analysis of shopping center customer behavior, compared to other methods?
- To what extent are the various assumptions about customer behavior and tenant placement, which can be found in the literature, validated in the analyzed shopping center?
- How can the assessment of retail category concentrations be improved?
- How can survey information on customer flows and coupling behavior be used by center management to improve profits?

**Variable Clumping Method and Mean-k-Nearest-Neighbor Method - Introducing two new approaches to retail concentration measurement to shopping center research**

- What are the deficiencies of existing methods for the assessment of retail concentration measurement?
- How can the *Variable Clumping Method* (*VCM*) be implemented in shopping center research and offer a probability-based analysis for assessing clustering on multiple spatial scales?
- How can the results of *VCM* be used for shopping center research and management?

**Do urban tourism hotspots affect Berlin housing rents?**

- What are the effects of urban tourism hotspots (attractions, hotels, restaurants and holiday flats) on rental prices in Berlin?
- How do building characteristics like age, space or location influence rents?
- How can spatial data on tourist amenities and urban infrastructure be integrated into the analysis?
- How can *generalized additive models* and especially *geoadditive models* improve the analysis?
- Is the observed increase in rents primarily a consequence of improved building characteristics?

## 1.3 Course of Analysis

This section provides an overview of the present status of the five papers.

**Energy efficiency: Behavioural effects of occupants and the role of refurbishment for European office buildings**

- **Authors:** Markus Surmann, Jens Hirsch
- **Submission to:** Pacific Rim Property Research Journal
- **First Submission:** 21.09.2015
- **Revised Submission:** 09.12.2015, 03.03.2016
- **Accepted for publication:** 17.03.2016

**Assessment of climatic risks for real estate**

- **Authors:** Jens Hirsch, Thomas Braun, Sven Bienert
- **Submission to:** Property Management
- **First Submission:** 28.01.2015

- **Revised Submission:** 30.04.2015
- **Accepted for publication:** 18.05.2015

**The analysis of customer density, tenant placement and coupling inside a shopping centre with GIS**

- **Authors:** Jens Hirsch, Matthias Segerer, Kurt Klein, Thomas Wiegelmann
- **Submission to:** Journal of Property Research
- **First Submission:** 20.04.2015
- **Revised Submission:** 31.08.2015, 21.10.2015, 15.12.2015
- **Accepted for publication:** 16.12.2015

**Variable Clumping Method and Mean-k-Nearest-Neighbor Method - Introducing two new approaches to retail concentration measurement to shopping center research**

- **Authors:** Jens Hirsch, Matthias Segerer
- **Submission to:** Journal of Geographical Systems
- **First Submission:** 10.02.2016

**Do urban tourism hotspots affect Berlin housing rents?**

- **Authors:** Philipp Schäfer, Jens Hirsch
- **Submission to:** International Journal of Housing Markets and Analysis
- **First Submission:** 04.05.2016

## 1.4 References


Bienert, S. (2014). *Extreme weather events and property values. Assessing new investment frameworks for the decades ahead*, Urban Land Institute, London.

Brown, S. (1992). Tenant mix, tenant placement and shopper behavior in a planned shopping centre, *The Service Industries Journal*, 12(3), 384–403.

Burke, M., Hsiang, S. M., Miguel E. (2015). Global non-linear effect of temperature on economic production, *Nature*, 527(7577), 235–239.

Carter, C. C., Haloupek, W. J. (2000). Spatial autocorrelation in a retail context, *International Real Estate Review*, 3(1), 34–48.

Carter, C. C., Vandell, K. D. (2005). Store location in shopping centers: Theory and estimates, *Journal of Real Estate Research*, 27(3), 237–265.

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., Way, R., Jacobs, P., Skuce, A. (2008). Quantifying the consensus on anthropogenic global warming in the scientific literature, *Environmental Research Letters,* 8(2).

Des Rosiers, F., Thériault, M., Lavoie, C. (2009). Retail concentration and shopping center rents: A comparison of two cities, *Journal of Real Estate Research*, 31(2), 165–207.

Eurostat (Ed., 2015). Eurostat Data Explorer, online: http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do (accessed: 01.04.2016).

Munich Re (Ed., 2015). Loss events worldwide 1980-2014, online: https://www.munichre.com/site/touch-naturalhazards/get/documents_E2080665585/mr/assetpool.shared/Documents/5_Touch/_NatCatService/Focus_analyses/1980-2014-Loss-events-worldwide.pdf (accessed: 01.04.2016).

Munich Re (Ed., 2016). Loss events worldwide 2015, online: http://www.munichre.com/site/corporate/get/documents_ E-397017904/mr/assetpool.shared/Documents/5_Touch/Natural%20Hazards/NatCatService/Annual%20Statistics/2015/2015_Torten_Ereignis_e.pdf (accessed: 01.04.2016).

Munich Re, GeoRisks Research, NatCatSERVICE (Ed., 2014). Internal supply of data on type, number and costs of natural disasters.

Nelson, R. L. (1958). The selection of retail locations, F. W. Dodge, New York, NY.

Stepper, M. (2014). *Stärkung der innerstädtischen Einzelhandelslagen vor dem Hintergrund des zunehmenden Online-Einkaufs*, in: Küpper, P., Levin-Keitel, M., Maus, F., Mueller, P., Reimann, S., Sondermann, M., Stock, K., Wiegand, T. (Eds.), Raumentwicklung 3.0: Gemeinsam die Zukunft der räumlichen Planung gestalten, 175–187, Akademie für Raumforschung und Landesplanung, Hannover.

Stern, N. (2006). *The Economics of Climate Change: The Stern Review*, Cambridge University Press, Cambridge.

UNEP United Nations Environment Programme (2009). UNEP Sustainable *Buildings* & Climate Initiative (Ed.), Paris.

World Economic Forum (2016). *Environmental Sustainability Principles for the Real Estate Industry*, Geneva.

Yuo, T. S.-T., Crosby, N., Lizieri, C., McCann, P. (2004). Tenant mix variety in regional shopping centres: Some UK empirical analyses, Working Papers in Real Estate & Planning, University of Reading.

# 2 Energy efficiency: Behavioural effects of occupants and the role of refurbishment for European office buildings

**Markus Surman** **Jens Hirsch**

**Abstract**

Energy consumption in office buildings is determined partly by fixed building characteristics, but also by the behaviour of occupants. Within the European Union, office buildings have become subject to more stringent energy efficiency regulation for new construction or extensive refurbishment, with the aim to reduce energy consumption and carbon emissions. The study determines the influence of physical building characteristics and occupant behaviour on energy consumption, and in particular, the role of refurbishment in different intensities on energy consumption is investigated. The data-set of the Green Rating Alliance is tested to provide evidence, by applying multiple regression models for energy consumption. The results highlight considerably increased energy consumption of single-tenant compared to multi-tenant office buildings. Very large office buildings consume significantly more energy per square metre that their smaller peers. A building's modelled water consumption turns out to be a good indicator for the actual energy consumption, emphasising the importance of assessing further sustainability measures. Overall, buildings of higher age turn out to be of lower energy consumption, pointing to additional appliances and equipment in more recent buildings, to provide better services and more comfort. In general, extensive refurbishment measures account for significant higher energy use, since the overall quality of the buildings is improved with additional appliances and equipment. Testing for the interaction effect between building age and refurbishment, the results demonstrate significantly lower additional energy consumption for buildings with more recent extensive refurbishment, compared to those with refurbishment several years ago. However, the results need to be considered with precaution against deriving firm conclusions due to the small sample size and some drawbacks in the applied data-set.

**Keywords:** Sustainable real estate, energy efficiency, carbon emissions, behavioural real estate, office buildings.

**Acknowledgments:** The authors are thankful to *Green Rating Alliance* (*GRA*) for provision of the dataset used in the study.

## 2.1 Introduction

In the *European Union* (*EU*), energy consumption by the building sector accounts for 40% of the total final energy use. Over the past decade, *EU* policy has required all member states to implement increased energy efficiency regulation for new construction. As a consequence, office buildings became subject to more stringent energy efficiency regulation for new construction or extensive refurbishment. The term 'refurbishment' is used within this paper referring to other terms used in the real estate industry, such as retrofit, redevelopment or revitalisation, usually defining major construction works affecting the thermal, technical and further energy consuming characteristics within the existing building structure. In a theoretical framework, these measures are expected to save energy, provided the technical facilities and their operability, as well as the behaviour of occupants, do not undermine the potential positive effects, or additional services with further equipment and additional energy consumption is applied.

The energy consumption of office buildings is determined mainly by core operations, which refer to physical building characteristics (heating, cooling, lighting, ventilation and elevators) and consumption from applied technical equipment, depending on occupant behaviour in the buildings. That is, besides the fixed building characteristics such as location, size, building fabric and age, energy consumption in office buildings also depends substantially on the behaviour of their occupants.

While more stringent energy efficiency regulation is intended to reduce carbon emissions from the existing building stock, the question arises as to whether this can be achieved essentially by focusing on the physical building characteristics in the context of technological progress. The controversial debate about the existence of a latent 'rebound effect', implying a negative behavioural response of occupants when confronted with a more energy- efficient building quality, pertains to the role of the occupants with their behaviour towards energy consumption. Since the rebound effect is attributed to occupant behaviour with additional use of the same services due to increased efficiency, it is rather difficult to identify this effect. For the existing building stock refurbishments are in many cases not realised in order to only increase the energy efficiency, but to increase or to provide further services, such as technical equipment that might be installed, in order to satisfy increased user requirements.

In this context, the study tries to investigate the relationship between actual energy consumption, physical building characteristics and behavioural attributes of occupants, subject to attempts to control for outdoor weather conditions and spatial heterogeneity.

The results provide evidence that refurbishments are associated with higher consumption in the tested office building portfolio, due to additional appliances and equipment. Testing a small sub-sample with refurbishment dedicated to energy efficiency and thermal building characteristics, no indication of a potential rebound or significant energy savings from these refurbishment measures are found. Although newer buildings are subject to more stringent energy efficiency regulation, our study reveals that newer office buildings do not necessarily consume less energy. This is found to be true only for most recently refurbished observations in the tested portfolio.

The remainder of this paper is as organised as follows. Section 2.2 provides the background to our study and considers some related research. In Section 2.3, we explain the characteristics of the used data-set and discuss the econometric methodology. Section 2.4 presents the results of the regression model, while Section 2.5 highlights some major conclusions and recommendations for further research.

## 2.2 Background and empirical framework

### 2.2.1 Background

Due to a significant greenhouse gas externality associated with energy consumption, the building sector has considerable potential in reducing global carbon emissions from the existing building stock. In the absence of any carbon pricing, an increment in energy consumption of commercial buildings has a negative impact on climate change.

Within the *EU*, the *European Performance of Buildings Directive* (*EPBD*) obliged all member states to implement increased energy efficiency regulation for new construction. In the course of the *EPBD*, Energy Performance Certificates (*EPC*) were introduced in 2002 and became mandatory by the year 2008. The regulation was implemented, based on a theoretical framework requiring the application of stricter building codes for new construction, and for existing structures undergoing complete or major refurbishment. This applies also for office buildings. The targeted reduction in energy consumption is based mostly on the potential offered by the physical building. However, little attention has been paid to the success factors of refurbishment in the context of energy savings.

Over the past decade, research insights into energy consumption and potential carbon emission savings in the commercial building sector of Europe have remained limited. This corresponds to experiences in the *US*, where Kahn, Kok, and Quigley (2014) found that research on commercial building energy consumption is still limited and most of it has been provided by engineers rather than economists. Research on the engineering dimension of energy efficiency of office buildings exhibits an extensive body of literature from the past decades, for example, in the related research projects of the *Association of European Renewable Energy Research Centres* (2016), *Fraunhofer ISE* (2015), *Fraunhofer IBP* (2009) or Post-occupancy Review of Buildings and their Engineering (CIBSE, 1997). However, a study by Guerra Santin et al. (2009) argues that empirical results, especially on occupants' influence in the commercial or office building sector of Europe, remain unsatisfactory.

The energy consumption of office buildings is determined by the combination and interaction of multiple factors. The physical building characteristics include location, building envelope (referring to building size, fabric and age) and technical equipment, such as heating, cooling, ventilation, lighting, elevators and *IT* equipment. The most relevant factors influencing energy consumption for heating and cooling are the thermal characteristics and related technical systems, the building type with regard to the surface to cubic volume ratio, occupant behaviour and the outdoor weather conditions. Since the largest energy consumption in office buildings is determined by heating, ventilation and air-conditions (*HVAC*), a significant intensification in the energy consumption, due to the expansion of *HVAC* systems in new buildings, was observed over the past decades. With the further expansion of new office space build, a growing trend of energy consumption is expected for the future (see Lombard et al., 2008).

Based on *UK* office buildings, Jenkins, Liu, and Peacock (2008) investigate that the energy consumption is primarily dominated by heating energy consumption. Their study assumes a more efficient office equipment and lighting in the future with lower levels of surplus heat production, but an increasing demand in heating energy for substitution. This will be mitigated to some extent by the temperature increase coming along with climate change.

Due to economies of scale in heating and cooling of office buildings, very large structures might behave differently to their smaller peers. Kahn et al. (2014) prove for significant higher energy consumption with

increase in building size. They assume that heating and cooling of large buildings requires additional equipment and energy loads to bridge large vertical distances in office towers, offsetting otherwise beneficial economies of scale.

The difference between actual energy consumption and engineering-predicted intrinsic energy consumption depends on the final construction with its installed technical systems and also on the utilisation of such systems, for example, in response to the indoor temperature set by occupants. However, the predicted intrinsic energy consumption is estimated on the basis of several determinants included in a modelled code baseline building, to indicate potential cost savings. Thus, the intrinsic assessment does not necessarily predict the future actual consumption since the prediction is applied as an engineering benchmark for relative energy performance, to allow for comparisons between buildings or implement stricter energy efficiency regulation in building codes (see New Buildings Institute, 2008). Torcellini et al. (2004) suggest that the deviation between actual consumption and predicted savings from intrinsic assessment is caused by higher than expected loads form occupants' behaviour and systems which do not perform together as designed.

The comparison between intrinsic predictions and actual consumption is even more difficult, since the intrinsic energy assessment might not predict all issues and variation in operational factors of energy consumption, such as 'plug loads'. These plug loads represent not the 'regulated loads' for basic building comfort, such as *HVAC* and lighting, but the 'unregulated' or process-related energy consumption, which is primarily driven by building equipment (elevators, computers, video-screens) and activity of building occupants. In the modelling of intrinsic energy for certification of *Leadership in Energy and Environmental Design* (*LEED*), a default of 25% of total 'baseline energy' was included. More importantly, previously completed *LEED* projects were not able to attempt any energy savings in the unregulated plug load category with a widely varying percentage of plug loads (see New Buildings Institute, 2008) and Scofield (2009) argues that evidence for lower energy from *LEED*-certification of office buildings has not been provided in a previous study.

Further key factors account for differences between the predicted intrinsic and actual energy consumption, such as differing occupancy hours and intensities, experimental – especially energy saving – technology does not perform as expected or a lack of knowledge, how to run the building most energy efficiently by facility managers and/or occupants (see Newsham et al., 2009). To sum it up, it remains reasonable that actual energy consumption is affected by multiple factors, whereas intrinsic energy assessment has limits in the applied determinants to predict actual energy consumption.

### 2.2.2 Behavioural effect of office occupants

Among the research of occupant behaviour and related effects on energy consumption exists only a small body of literature and the most of it is dedicated to modelling tools to simulate the influence of occupant behaviour and how occupants interact with building equipment and plug loads.

Besides the fixed influence of applied technical equipment in buildings, occupant behaviour concerning *HVAC* is highly dynamic and depends not only on outdoor climatic conditions, but also on the type of *HVAC* equipment and occupant experiences with them. Individual heating or cooling systems, instead of a centralised control system, allow for a varying usage between different (parts in) office buildings.

Technical equipment and plug loads in office buildings are prone to be significantly influenced by the behaviour of occupants in different appliances and intensities. Another influence factor is whether the

building is rented out to a single tenant or to several. The allocated office space per occupant implies an important effect as well as the overall occupancy rate, indicating business cycle effects. Kahn et al. (2014) found that a 1% higher occupancy rate increases electricity consumption by 2.6% in office buildings. Depending on the specific industry and the related technical equipment, such as IT, occupants' activities and behavioural patterns result in a different intensity of energy consumption.

Moreover, the individual awareness and behavioural attitude of occupants towards energy consumption and potential energy (cost) savings is assumed, to play an overall important role in the dynamic dimension of energy consumption (see Bloom et al., 2011). Experience from the *US* demonstrates that the presence of a building engineer significantly lowers consumption, compared to buildings without an engineer (see Kahn et al., 2014). For Australian office properties, Gabe (2014) found that a frequent site energy consumption auditing is a potential strategy to reduce energy consumption and mitigate greenhouse gas emissions. As part of a so-called green management strategy, the repetitive auditing experiences to be a successful approach for motivating owners to invest in energy efficiency technologies. For a European portfolio of corporate real estate assets from the wholesale and hypermarket sector, Surmann et al. (2016) found evidence that a centralised corporate energy management contributes to recurring energy consumption reductions and thus energy measures for corporate assets provide leverage towards a more efficient corporate environmental performance.

Research work of Kavulya and Becerik-Gerber (2012) analysed occupant behaviour in an office environment and interaction of occupants with energy consuming equipment in visual observations with tracking of daily activities on commonly used office appliances. The results estimate an energy saving potential of up to 38%, if occupants switch of appliances not in use, due to higher awareness between consumption and occupant usage data. The study argues that energy awareness plays a key role to modify the behaviour of occupants towards reduced energy consumption.

### 2.2.3 Refurbishment and rebound effect

The term rebound effect refers to a situation in which the actual energy savings from an innovation are lower than those expected from improved efficiency, due to more extensive – rebound – consumption by users, either in the form of more hours of use or a higher quality of energy service (see Herring and Roy, 2007). Experience from the automobile industry shows that a reduction in fuel consumption was achieved, while the safety and comfort attributes of cars had been enhanced remarkably (see Knittel, 2012). With regard to cars, the term rebound effect means that a more fuel efficient car will lead to more kilometres travelled (see Gillingham et al., 2013).

Similar observations were expected from the commercial building sector in the past, offsetting increased energy efficiency to some extent. The effect has been investigated and described in an early study for commercial buildings by Greening, Greene, and Difiglio (2000). They conclude that the range of estimates for the size of the rebound effect is very low to moderate. Based on a review of studies for gasoline and electricity consumption, Gillingham et al. (2013) and Gillingham et al. (2016) argue that the behavioural response of users offsets between 5% and up to 30% of intended energy savings. They conclude that the rebound effect is rather small and therefore no excuse for inaction in the economy.

For commercial real estate, it is difficult to distinguish between the rebound effect and the 'principle of additionality', in which higher energy efficiency is realised with the provision of increased or additional services and comfort, provided with less energy consumption than they would otherwise have. This principle of additionality usually comes along in the course of refurbishments or in the development of new

buildings, when additional or new technical equipment is installed, in order to satisfy increased user requirements. The principle is observed for commercial buildings of lower age or recent refurbishment with higher energy consumption compared to their older peers. Recent research results from the *US* reveal that both younger office buildings and those of higher quality are in fact responsible for higher electricity consumption (see Kahn et al., 2014).

Kahn et al. (2014) state that energy consumption and building quality are complements – not substitutes. Even when technological progress reduces the theoretical energy demand from *HVAC* and lighting, the increase in quality attributes, such as a more attractive lobby and office space, more elevators and individual adaption of comfort temperature by occupants, may actually increase energy consumption. Hence, the replacement of older structures by new buildings or at least extensive refurbishment is likely to increase the energy consumption of the durable building stock.

Results for a refurbishment variable included in the work of Kahn et al. (2014) have documented that refurbished buildings feature a higher energy consumption of 19%, compared to similar-sized buildings without refurbishment. Besides a potential, but expected to be small, rebound effect when improved building quality provides better HVAC and lighting systems which may induce greater energy use, the additional services employed in the course of refurbishment account for the increase in consumption for the most part.

Furthermore, refurbishment is not associated only with energy-saving measures of the technical equipment in a building, but often involves a replacement or enlargement of the technical infrastructure, especially lighting, *HVAC* and *IT*, which might be associated with additional energy consumption. A survey by Kok et al. (2012) found only 14% of refurbishments with improvements solely dedicated to sustainability, whereas refurbishment was carried out to improve the overall quality of the buildings. In the context of the necessity to update otherwise obsolete buildings (see Baum and McElhinney 1997, Baum and Turner, 2004) also energy efficiency improvements were considered for the technical equipment replaced, but moreover the building quality standard as a whole is enhanced. This corresponds to the observation of Chegut et al. (in press) that although sustainable construction is gaining market share, new construction and building refurbishments are still mostly conventional.

With regard to the discussed empirical findings, our study is intended to investigate the relationship between actual energy consumption, physical building characteristics and occupant attributes. With increasing building size up to a certain point, we expect lower energy consumption by trend for office buildings, due to economies of scale in heating and cooling. For very large office buildings (office towers), we expect higher energy consumption while economies of scale are offset by higher energy loads to bridge large vertical distances in office towers. Since energy efficiency regulation within the *EU* has become more rigorous for new construction or extensive refurbishment, we test whether higher energy efficiency is achieved for younger office buildings. The effect of refurbishment is of special interest in this study. On the one hand, major refurbishment is expected to consider higher energy efficiency standards, thus allowing for improvements of the thermo-physical quality of the buildings and potential energy savings. On the other hand, research results prove for additional energy consumption in the course of refurbishment, due to the provision of additional services and new technical equipment employed.

## 2.3 Dataset

In order to answer our research questions, we use the data-set of the *Green Rating Alliance* (*GRA*), which provides physical office building characteristics and occupant attributes in detail at the building level. The

data-set includes two main sources of energy consumption; first, the actual metered energy consumption and second, the intrinsic energy consumption as result of the Green Rating auditing. However, the data-set of *GRA* contains actual energy consumption and intrinsic assessment metered only once at a certain point of time in the years from 2008 to 2012 for issuing the Green Rating audit. Therefore, the data-set is not covering a panel of observations over time, which is a drawback for the analysis.

The intrinsic energy measure of *GRA* is based on an individual assessment of the physical building characteristics, with an estimation of the thermal qualities of different construction and fit-out elements, inherent in each single building. The calculation of intrinsic consumption is modelled under standard – optimised – conditions of the building use with assumptions of occupant behaviour concerning schedules and temperature set points. While this fixed standard model is equal for each building in the assessment, the intrinsic energy should be seen as a measure of potential energy consumption, without taking into account the impact of occupant behaviour differing among buildings. Since *GRA*'s intrinsic benchmark is modelling under optimised conditions of the building use and does not account for plug loads from the occupants, we expect much lower intrinsic figures than actual metered consumption.

The data-set also contains actual and intrinsic measures concerning the buildings' annual water consumption. Parallel as for energy consumption, actual measures are derived from metered consumption and intrinsic measures from modelling based on standard assumptions. In regard to the relationship between energy and water consumption, office buildings designed with higher energy efficiency standards or stricter building codes might also have higher standards in water efficiency for lower actual consumption. Investors committed for investment in sustainable real estate consider metered water consumption besides energy consumption. Among other institutional investors, *Bouwfonds Investment Management* (2013) states commitment, to contribute to a reduction in water intensity of the real estate they manage. We hypothesise a correlation between energy and water consumption, due to a more efficient building design and higher efficiency requirements from increased regulation, as well as from the social responsible investment strategies of institutional investors (see Cajias and Bienert, 2011; Cajias et al., 2011; Kerscher and Schäfers, 2015).

Therefore, we use intrinsic water consumption measures from *GRA* to estimate actual energy consumption, because the intrinsic consumption is based directly on building characteristics and not influenced by occupant behaviour, thereby avoiding potential bias.

To control for the attributes of outdoor weather conditions and temperature, the *heating degree days* (*HDD*) and *cooling degree days* (*CDD*) of the respective auditing year were used. The heating and *CDD* were calculated on a basis of 65°F (18.3°C), obtained from the database of *Weather Underground* (2015). While the auditing process of *GRA* with actual measurement is assessed over a time period of at least 8 months (normally 12), we applied the *HDD* and *CDD* with respect to the year in which more than 4 months of the auditing period were carried out. Our total sample comprises 288 observations, combining the *GRA* sample with the *HDD* and *CDD* in a period from 2008 to 2012.

In the context of our theoretical considerations on the role of physical building characteristics, building age and size, as well as ceiling height and heating production type and the intensity of refurbishment are of a major research interest. To investigate the influence of occupants, we focus on the attributes of building area per office occupant, and differentiate between single and multi-tenant-occupied buildings. We expect a significant difference between buildings rented on single-tenant basis, compared to those on a multi-

tenant basis. Our supposition is that multi-tenant buildings face more decisions regarding the heating, cooling and lighting of the office space in question, thus resulting to higher consumption.

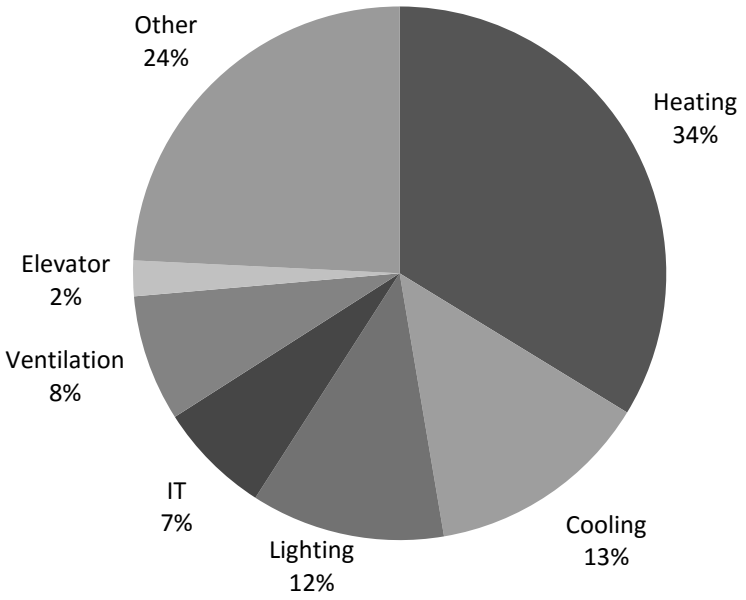
The sub-categories of the total actual energy consumption applicable from the data-set, enable distinguishing between each energy-consuming sub-category. Table 1 includes some descriptive statistics of the applied metric response and explanatory variables:

**Table 1: Descriptive statistics of applied metric attributes**

| Descriptive Statistics | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Actual energy consumption in kWh/sqm/a | 73.5 | 175.7 | 235.2 | 254.8 | 317.2 | 696.5 |
| Intrinsic energy consumption in kWh/sqm/a | 46.2 | 104.3 | 129.0 | 138.2 | 162.1 | 368.7 |
| No. of heating degree days in year of Green Rating Audit | 1,317 | 2,429 | 2,652 | 2,757 | 3,009 | 3,941 |
| No. of cooling degree days in year of Green Rating Audit | 0 | 123 | 163 | 179 | 183 | 734 |
| Intrinsic water consumption in cbm/sqm/a | 0.094 | 0.266 | 0.333 | 0.352 | 0.432 | 1.060 |
| Actual energy consumption heating in kWh/sqm/a | 6.1 | 49.1 | 75.5 | 83.5 | 101.6 | 365.7 |
| Actual energy consumption cooling in kWh/sqm/a | 0.0 | 12.1 | 24.5 | 34.6 | 41.6 | 205.0 |
| Actual energy consumption lighting in kWh/sqm/a | 1.0 | 17.5 | 24.2 | 27.5 | 35.3 | 115.6 |
| Actual energy consumption IT in kWh/sqm/a | 1.1 | 8.3 | 12.8 | 15.9 | 20.7 | 77.1 |
| Actual energy consumption ventilation in kWh/sqm/a | 0.0 | 8.4 | 14.6 | 19.8 | 25.1 | 144.7 |
| Actual energy consumption elevator in kWh/sqm/a | 0.0 | 2.3 | 3.9 | 5.3 | 6.3 | 61.1 |
| Actual energy consumption other in kWh/sqm/a | 0.0 | 24.5 | 50.2 | 68.7 | 93.1 | 503.0 |
| Building age (economic) | 0.0 | 3.0 | 9.0 | 11.7 | 19.0 | 50.0 |
| Building area in sqm | 1,340 | 5,596 | 10,040 | 13,960 | 17,979 | 108,070 |
| Building area in sqm per occupant | 5.2 | 17.0 | 21.8 | 26.2 | 30.7 | 147.1 |
| Ceiling height in meters | 2.3 | 2.6 | 2.7 | 2.8 | 2.9 | 4.3 |

Comparing the actual with the intrinsic energy consumption, a large gap is obvious at first glance. Since the intrinsic consumption is a measure in relation to the physical building characteristics and a standard factor for occupant influence, modelling under optimised conditions of the building use, the large gap meets our expectations.

**Figure 1: Sub-categories' share of total actual energy consumption**




When looking at the share of sub-categories to the total energy consumption, it is evident that heating, cooling and ventilation account for more than 55% of the total actual consumption. The sub-category

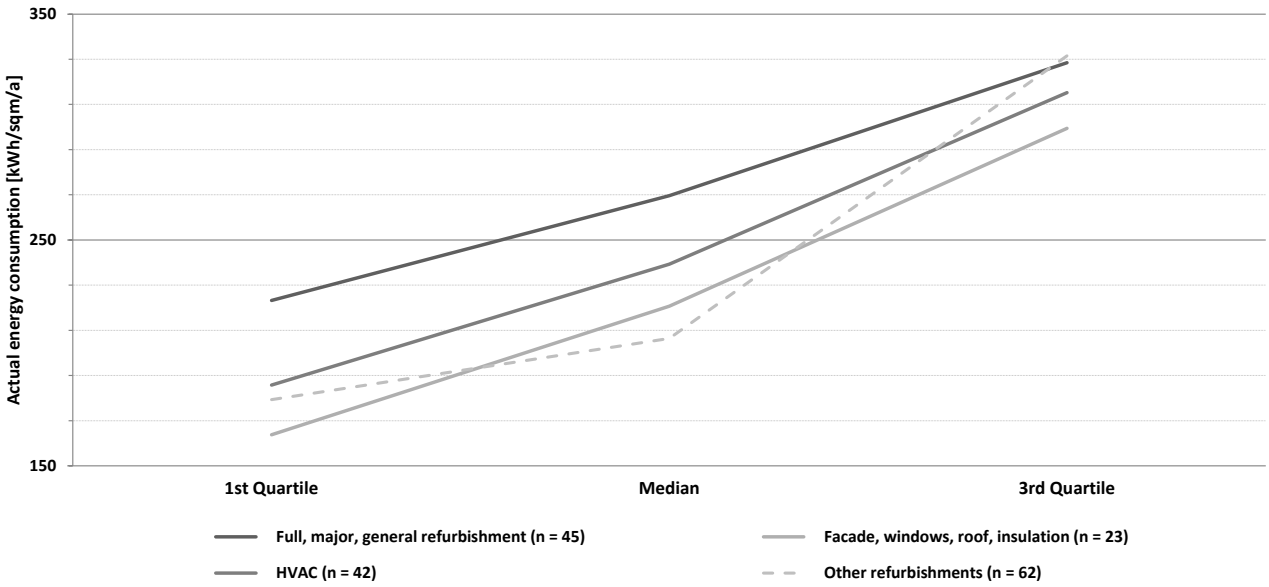
'other' accounts for a share of 24% of the total consumption, including consumption e.g. from (underground) car park, canteen and outside lighting. However, the share of these categories summarised under 'other' was not applicable from the *GRA* data-set, what points to a limitation when explaining actual energy consumption.

The economic building age was derived from the (historical) construction year under consideration of complete or extensive (full / major / general) refurbishment in the past and yields an average of 12 years. This result demonstrates the importance of the information from the data-set regarding refurbishment.

**Table 2: Actual energy consumption and refurbishment sub-samples**

| Refurbishment and Actual energy consumption [kWh/sqm/a] | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| Total Sample (n = 288) | 73.4 | 175.7 | 235.2 | 254.8 | 317.2 | 696.5 |
| Refurbishment = no (n = 115) | 86.3 | 169.6 | 223.2 | 240.4 | 296.1 | 696.5 |
| Refurbishment = yes (n = 173) | 73.4 | 181.6 | 243.1 | 264.4 | 328.5 | 685.5 |
| Refurbishment <= 5 years (n = 109) | 73.5 | 175.9 | 238.4 | 262.4 | 328.5 | 685.5 |
| Refurbishment > 5 years & <= 15 years (n = 50) | 111.6 | 188.7 | 240.4 | 256.1 | 294.3 | 615.2 |
| Refurbishment > 15 years (n = 14) | 134.4 | 190.2 | 251.2 | 310.2 | 444.8 | 503.3 |
| Full, major, general refurbishment (n = 45) | 104.0 | 223.2 | 269.6 | 276.1 | 328.5 | 478.6 |
| Façade, windows, roof, insulation (n = 23) | 91.8 | 163.8 | 220.7 | 243.9 | 299.4 | 538.8 |
| HVAC (n = 42) | 108.6 | 185.8 | 239.3 | 261.6 | 315.2 | 609.7 |
| Other refurbishments (n = 62) | 104.0 | 179.3 | 206.4 | 263.7 | 331.5 | 685.5 |

**Figure 2: Actual energy consumption and type of refurbishment**




At first glance, we find no verification that refurbishment has a positive impact on energy efficiency in such a way, as to decrease actual energy consumption of the tested office buildings (see Table 2). Mean and median values of actual consumption of refurbished buildings are slightly above the levels of non-refurbished buildings. In comparison to the total sample, the observations attributed to having undergone a refurbishment in the past (n = 173) suggest a 3.6% higher energy consumption on average which points to the principle of additionality. The sub-sample for buildings refurbished more than 15 years ago indicates significantly higher actual energy consumption, especially concerning the mean and 75%-percentile, when compared with more recently refurbished buildings (see Figure 2). It can be concluded that refurbishments in the last 15 years came with a stronger focus on energy efficiency. Actually, the share of refurbishments

with a focus on façade, windows, roof, insulation or *HVAC* is much higher for recently refurbished buildings (≤15 years: 23.2%) than for other buildings (>15 years: 4.5%).

A small sub-sample of 23 observations attributed with energy efficiency refurbishment concerning façade / windows / roof / insulation, which is expected to reduce energy consumption, shows the lowest actual energy consumption, as regards first, second and third quartile and the mean value as well as compared to the total sample. The result indicates a conservation potential of around 4.5% on average and approximately 6.6% for the median value, due to the refurbishment of thermal building characteristics. In contrast to the slight decrease in actual consumption in relation to the refurbishment of only thermal building characteristics, we found the actual consumption of 45 observations attributed with full / major / general refurbishment with a 13.2% higher consumption on average. This sub-sample even exceeds the average per square metre consumption of the total sample by 7.7%. Buildings with extensive (full/major/general) refurbishment also show higher energy consumption than buildings with refurbishments concerning *HVAC*-equipment. We assume that for extensive refurbishments, the potential positive effects of renewing thermal building characteristics or *HVAC* equipment might be counteracted by other changes in the building resulting into higher energy consumption. At first glance, the result proves for the principle of additionality as effect of extensive refurbishment measures.

To verify these preliminary results while accounting for other effects, we include refurbishment details as explanatory variables in our regression analysis.

**Table 3: Correlation matrix of metric attributes**

| **Correlation Matrix** | | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LN(actual energy consumption in kWh/sqm/a) | (1) | 1 | | | | | | | | | |
| LN(intrinsic energy consumption in kWh/sqm/a) | (2) | 0.544 | 1 | | | | | | | | |
| Relative spread between intrinsic and actual energy consumption | (3) | 0.377 | -0.289 | 1 | | | | | | | |
| No. of heating degree days in year of Green Rating audit (HDD) | (4) | 0.084 | 0.225 | -0.080 | 1 | | | | | | |
| No. of cooling degree days in year of Green Rating audit (CDD) | (5) | -0.200 | -0.254 | 0.031 | -0.556 | 1 | | | | | |
| Intrinsic water consumption in cbm/sqm/a | (6) | 0.233 | 0.306 | -0.048 | 0.004 | -0.072 | 1 | | | | |
| Building age (economic) | (7) | -0.056 | 0.020 | -0.004 | -0.089 | 0.089 | 0.057 | 1 | | | |
| LN(building area in sqm) | (8) | 0.126 | -0.043 | 0.054 | 0.073 | -0.044 | -0.118 | -0.054 | 1 | | |
| LN(building area in sqm per occupant) | (9) | -0.143 | -0.019 | -0.139 | 0.039 | -0.061 | 0.127 | -0.029 | 0.023 | 1 | |
| Ceiling height in meters | (10) | 0.076 | 0.125 | -0.040 | 0.112 | -0.047 | -0.012 | -0.174 | 0.033 | 0.082 | 1 |

The correlation matrix for metric attributes (Table 3) shows an orthogonal linear relationship between the response and explanatory variables. Corresponding to our expectations, a positive bivariate relationship between intrinsic and actual energy consumption is observable. *HDD* (*CDD*) demonstrate a positive (negative) bidirectional relationship with actual and intrinsic consumption. Somewhat surprisingly, the energy consumption decreases with an increase in *CDD*. Anyway, this result is only based on correlation and needs to be further examined by a regression analysis, taking into account all other possible explanatory factors. The intrinsic water consumption shows a positive bivariate correlation with actual energy consumption.

The negative relationship between building age and actual consumption is at first glance interesting, when observations of older buildings turn out to be less energy consuming. The building size is expected to be inversely related to energy consumption (per square metre), due to economies of scale in larger buildings.

However, this does not appear to be true for actual consumption. The building area allocated per occupant has been calculated from the building area and the total number of occupants, both obtained from *GRA* data. The negative bivariate correlation between the building area allocated per occupant might be the result of missing information about vacancy as more space per occupant can mean: (a) a higher vacancy rate and (b) larger offices or common spaces. The two possible explanations cannot be distinguished from one another with the given *GRA* data.

## 2.4 Econometric approach

The energy consumption of office buildings can be explained with a function of the consumption from core operations (*HVAC*) in relation to the physical building and additional consumption from appliances with installed equipment and plug loads both used by the occupants. In order to determine these effects on the response variable, we have to control for all other factors affecting energy consumption. To address this issue within a multiple regression model, the dependent variable is decomposed into the implicit contribution of the available building characteristics and occupant attributes, while controlling for the outdoor weather conditions and effects from spatial heterogeneity.

The general regression model via ordinary least squares is described in Equation 1, with Y as the response variable, X as a vector containing the explanatory variables, β as unknown parameters and ε as error term.

$$Y_{\mathrm{i}} = \alpha_{\mathrm{i}} + \beta X_{\mathrm{i}} + \varepsilon_{\mathrm{i}} \qquad (1)$$

In our approach, the regression model is used for actual energy consumption metered in kWh/sqm/a. The vector containing the explanatory variables includes the physical building and occupant attributes, as well as the *HDD* and *CDD* controlling for the local outdoor climate weather conditions and effects from spatial heterogeneity.

For the regression of actual energy consumption, the response variable is transformed logarithmically (see Malpezzi, 2003). This procedure allows for the interpretation of the estimated effects as elasticities if both sides are logarithmically transformed, or semi-elasticities if the explanatory variable enters the equation in absolute values. Furthermore, strictly positive metric variables are transformed logarithmically when estimating a log-linear function in Equation (2):

$$\begin{aligned}[\mathrm{LN}(\text{actual E cons.})_i] = \alpha + \beta_1 HDD_i + \beta_2 CDD_i + \beta_3 \begin{bmatrix} \text{building age class}_i \\ \text{building area class}_i \\ \mathrm{LN}(\text{area/occupant})_i \\ \text{intrinsic water cons.}_i \end{bmatrix} + \\ \beta_4\ \text{single tenant binary}_i + \beta_5\ \text{electric heating binary} + \beta_6\ \text{refurb. binary}_i + \\ [\text{country}_i \kappa] + \varepsilon_i \end{aligned} \qquad (2)$$

where *(actual E cons.)*$_i$: represents the actual energy consumption of building *i*, *(area/occupant)*$_i$: represents the allocated area per occupant, *intrinsic water cons.*$_i$ represents the intrinsic water consumption; *single tenant binary*$_i$: represents single or multi-tenant use, *electric heating binary*$_i$: represents electric heat production, *refurb. binary*$_i$: represents full / major / general refurbishment, *country*$_i$: represents location (country) and $\varepsilon_i$: represents an *iid* error term.

The explanatory variables *HDD* and *CDD* control for the local outdoor weather conditions in the relevant year of the Green Rating audit of the observations (building i).

The *building age* is included to test for differences between newer and older buildings. This was considered under economic considerations, reflecting total or major refurbishment with improved physical characteristics of the buildings, thus expressing a proxy for depreciation. This economic building age was classified in three groups (0–10, 11–20, >20). The *building area* is introduced to the regression model in terms of dummy variables representing the building's belonging to one of five quantiles. The building area allocated per occupant in square metres is entered logarithmically into the model with *LN(area/occupant)*.

The Green Rating audit data also contain information about the buildings' sustainability characteristics besides energy consumption, e.g. *Water consumption*, waste and carbon emissions, metrically scaled in units per square metre and year. Carbon emissions are highly correlated with energy consumption and therefore omitted for the regression. If one of the non-energetic sustainability ratings has significant explanatory power for a building's energy consumption, sustainability characteristics might become more relevant for investors, because they can help to identify energy-efficient buildings. In this regard, we use the intrinsic values for water consumption, to assess the influence on energy consumption. The actual values are not considered, due to expected bias by the specific occupant behaviour. Furthermore, a dummy variable for *electric heating* enters the equation to control for energy consumption related to different technical systems in the buildings.

With regard to the occupant attributes, the *single tenant binary* distinguishes between a single or multi-tenant use of the building premises.

In order to provide evidence for the effect of refurbishment on energy consumption, another dummy variable is introduced, to estimate the effect of a *full, major or general refurbishment*. Since we have few observations in the data-set with refurbishment dedicated to energy efficiency and thermal building characteristics, we use the attributes *façade / windows / roof / insulation* instead of the binary for extensive refurbishment in a second specification of the regression model. For the small sub-sample of 23 observations, a positive effect in the regression of energy consumption might point to a potential rebound effect with no energy savings resulting from the refurbishment. A negative coefficient could indicate energy savings due to the refurbishment measures.

A matrix of *country* dummies was considered to control for spatial heterogeneity, e.g. energy efficiency regulation with regard to local building codes and different pricing of energy between the 14 European countries. Due to the introduced *HDD* and *CDD* with reference to the location-based outdoor weather conditions, we do not append additional location dummies, to control for spatial heterogeneity, so as to avoid any selection bias.

Since the assumption of linearity in the effects of regression models often seems to be too restrictive in a real estate context (see e.g. Brunauer et al., 2012; Brunauer et al., 2010; Mason and Quigley, 1996; Pace, 1998; Parmeter et al., 2007), it seems appropriate to use more flexible non-and semi-parametric regression models. For example, the effect of building age is known to be nonlinear (for instance, Fahrmeir and Tutz, 2001). We consider *generalised additive models* (*GAM*), as described in Wood (2006), to discover nonlinear effects for the continuous covariates. Applying *GAM* provides the advantage, to express the nonlinear effects in the relationship between response and explanatory variables in visualised nonlinear regression splines.

To control for nonlinearity in the effects of our regression model, we replace the linear effects $\beta_j x_{ij}$ of the continuous covariates with possibly nonlinear functions $f_j\left(x_{ij}\right)$ in equation (3):

$$[\mathrm{LN}(\text{actual E cons.})_i] = \alpha + \beta_1 \mathrm{HDD}_i + f_1\,(\mathrm{CDD}_i) + f_2\,(\text{building age}_i) + f_3\,(\text{building area}_i) + f_4\,(\mathrm{LN}(\text{area/occupant})_i) + f_5\,(\text{intrinsic water cons.}_i) + f_6\,(\text{refurb. binary}_i * \text{building age}_i) + \beta_2\,\text{single tenant binary}_i + \beta_3\,\text{electric heating binary}_i + \beta_4\,\text{refurb. binary}_i + [\text{country}_i \kappa] + \varepsilon_i \quad (3)$$

The linear effect for *HDD* is not replaced with a nonlinear function for technical reasons. It is common practice in regression modelling to introduce combined variables for important effects to estimate the interaction between the both variables. To further investigate the effect of refurbishment on energy consumption, the building age of those observations attributed with refurbishment from the past is introduced with a nonlinear function. Besides the separate main effect of full / major / general refurbishment, the interaction effect of building age for observations with a refurbishment in the past is included in the regression. This procedure allows to estimate and display the effect, to be interpreted as the additional energy consumption of observations undergone a refurbishment by trend.

## 2.5 Research results

### 2.5.1 Results for actual energy consumption from equation (2)

The results of our log-linear regression model for actual energy consumption are summarised in Table 4. The results for the regression with extensive refurbishment attributes full/major/general as explanatory variable are shown in column 1. The model specification in column 2 adds the observations for refurbishment dedicated to energy efficiency and thermal characteristics with façade / windows / roof / insulation instead of the extensive regression observations. In regard to the data-set employed, the adjusted $R^2$-value of both models appears to be low. Therefore, the results are to be considered with precaution.

For the tested portfolio, the influence of outdoor weather conditions on actual energy consumptions seems to be very low, while the coefficients of *HDD* and *CDD* are not of any significance.

Turning to further explanatory variables measuring the influence of physical building characteristics, we found the classified economic building age to have virtually no explanatory power and lacking greater significance. Only in the model specification including refurbishment with regard to energy efficiency and thermal characteristics in column (2) displays a low significant effect of those buildings with the highest economic building age (21–50 years). The actual energy consumption per square metre is 12% lower than in the group of youngest buildings (0–10 years). For the interpretation of the coefficients for binary variables in a semi-logarithmic regression, the percentage effect is calculated as anti-logarithm of the estimated coefficients with $\left((exp(\beta_j) - 1) * 100\right)$ with regard to the omitted reference variable (see Halvorsen and Palmquist, 1980; see Hardy, 1993).

While the economic building age was derived under consideration of the (historic) construction year and the applicable refurbishment year, as well as the intensity of refurbishment, we also run the regression

model with the (historic) construction year and found no significant effect, again. This is a remarkable finding, due to the overall expectation of an impact from the increased energy efficiency regulation in *EU*-member states over the past decade. Subject to potential limitations in the data-set with only 288 observations, this result would suggest that stricter building codes and construction standards seem to emerge without exerting a significant impact on the conservation of actual energy consumption.

Regarding the building area, the expectation of less consumption in heating and cooling does not appear for the tested portfolio. On the contrary, the quantile with the largest buildings even shows a significant increase in energy consumption between 15 and 16%, compared to the quantile with the smallest buildings. The data-set might include some high-rise office towers, contained in the quantile with the largest buildings and attributed with much higher consumption per square metre.

**Table 4: Regression results of log-linear model on actual energy consumption as response variable from equation (2)**

| **Response variable: LN(actual energy consumption in kWh/sqm/a)** | | |
|---|---|---|
| Explanatory variables coefficient (t-values) | 1 | 2 |
| Intercept | 5.700<br>(17.069)*** | 5.702<br>(17.141)*** |
| Number of HDD in audit year/100 | -0.011<br>(-1.534) | -0.008<br>(-1.202) |
| Number of CDD in audit year/100 | -0.055<br>(-1.674) | -0.060<br>(-1.761) |
| Economic building age 11-20 years | 0.063<br>(1.273) | 0.053<br>(1.077) |
| Economic building age > 20 years | -0.101<br>(-1.654) | -0.128<br>(-2.141)* |
| Building area second quantile | 0.010<br>(0.189) | 0.008<br>(0.143) |
| Building area third quantile | -0.015<br>(-0.206) | -0.030<br>(-0.435) |
| Building area fourth quantile | 0.060<br>(0.887) | 0.035<br>(0.530) |
| Building area fith quantile | 0.154<br>(2.365)* | 0.141<br>(2.165)* |
| Single-tenant | 0.142<br>(2.909)** | 0.147<br>(2.959)** |
| LN(sqm building area/occupant) | -0.127<br>(-2.409)* | -0.122<br>(-2.307)* |
| Intrinsic water consumption (cbm/sqm/a) | 0.482<br>(2.882)** | 0.483<br>(2.789)** |
| Electric heating | -0.192<br>(-3.665)*** | -0.200<br>(-3.791)*** |
| Full/major/general refurbishment | 0.139<br>(2.344)* | |
| Façade/windows/roof/insulation refurbishment | | -0.014<br>(-0.228) |
| Countries (n) | 14 | 14 |
| $R^2$ | 32.47 | 31.78 |
| Adjusted $R^2$ | 25.64 | 24.88 |

Significance: *** 1%, ** 5%, * 10%

The building area allocated per office occupant has a significant effect on energy consumption, indicating a decreased consumption for an increasing area per occupant. At first, this points to the relationship of higher vacancy in office buildings being associated with lower per square metre consumption. Furthermore, the evidence suggests that occupants with more space allocated, but constant equipment and plug loads (ceteris paribus) consume less per square metre. In other words: More occupants on the same office area will increase energy use per square metre, which is in line with the literature review in Section 2.2. Further explanatory power is expected from the regression with nonlinear effects and illustration in regression splines following Equation (3).

The single tenant binary provides significant results. The coefficients demonstrate that single-tenant office buildings have significantly higher energy consumption than multi-tenant- occupied ones. With regard to the omitted multi-tenant reference category, the dummy variable explains a higher actual consumption of 15% (both column 1 and 2). Our expectation was that multi-tenant buildings would have higher energy consumption, because of somewhat contradictory decisions when running the building. A conclusion might be that a single tenant intends more to heat, cool, ventilate and light a building centrally as a whole, not differentiating (even when possible) between single building parts or floors, e.g. cooling of upper floors only. According to this interpretation, within a multi-tenant building, each tenant might behave in a more decentralised manner specifically for the smaller occupied part of the building. The more, a large single tenant might potentially consider energy prices with minor importance when referred to business turnover and allocated headcount cost.

The variable for intrinsic water consumption per square metre and year shows significance as explanatory variable for energy consumption. An increase in water consumption of 100 litres per square metre and year comes along with an approx. 6% increase in actual energy consumption. A water-efficient building infrastructure seems to be an indicator for a better energy performance by trend.

The electric production for heating has a highly significant negative effect on energy consumption. Under control of all other factors, energy consumption per square metre is more than 20% lower for electric heating in comparison to district heating network or boilers, thus indicating higher energy efficiency.

Turning to the refurbishment details introduced in our econometric approach, we obtain a significant coefficient for the aggregated dummy, containing full/major/general refurbishment, with reference to omitted observations without any refurbishment. The result reveals a positive impact on actual energy consumption for observations attributed with an extensive refurbishment. Compared to buildings without refurbishment, extensive refurbishment is attributed to have a higher energy consumption of 15%. The result proves for the principle of additionality being associated with refurbishment for the tested portfolio. This suggests that refurbishment was carried out to provide increased or additional services and comfort from appliances with additional or new equipment to satisfy user requirements.

Turning to the model specification with the small sub-sample for refurbishment dedicated to energy efficiency and thermal building characteristics, we find no significant impact of refurbishment associated with thermal qualities, contained in façade / windows / roof / insulation. The coefficient is estimated with a negative value, which is intuitive to our expectation, but insignificant.

### 2.5.2 Results for actual energy consumption from equation (3)

The introduction of *GAM* and visualisation with smoothed curves in regression splines is considered to control for nonlinearity in the effects of the regression. Table 5 provides the parametric coefficients for the linear effects following Equation (3).

While lacking any significance, the continuous covariate for *HDD* was not visualised with a regression spline. For the single tenant binary, the results show again significantly higher energy consumption for single-tenant-occupied buildings. The coefficient for electric heating proves for significantly higher energy efficiency, again.

Interestingly, the introduced dummy variable for full / major / general refurbishment shows a positive coefficient, but is lacking significance. When this separate main effect is remaining without significance, we expect the interaction effect of building age for observations with a refurbishment in the past to explain the additional energy consumption of buildings with refurbishment in relation to their age.

The model with application of *GAM* in Equation (3) reveals a higher explanatory power with adjusted $R^2$ above 35%, compared to the linear model specification from Equation (2). However, the unexplained variation in the models remains of a considerable level.

The further covariate effects from Equation (3) are illustrated in the regression splines of Figure 3. In each graph, the y-axes can be approximately analysed as the percentage effect on energy consumption. The value for the *estimated degrees of freedom* (*edf*) higher than 1.0 displays a nonlinear function in the relationship. Within the splines, the continuous black lines are the expected effects and the grey areas are point-wise 95% confidence intervals.

**Table 5: Parametric coefficients of log-linear model on actual energy consumption as response variable from equation (3)**

| **Response variable: LN(actual energy consumption in kWh/sqm/a)** | | |
|---|---|---|
| Explanatory variables coefficient (*t*-Values) | | |
| Intercept | 5.120 | (15.881)*** |
| Number of HDD in audit year/100 | -0.006 | (-0.828) |
| Single-tenant | 0.100 | (2.050)* |
| Electric heating | -0.219 | (-4.009)*** |
| Full/major/general refurbishment | 0.091 | (1.443) |
| Countries (n) | 14 | |
| $R^2$ | 44.50 | |
| Adjusted $R^2$ | 36.01 | |

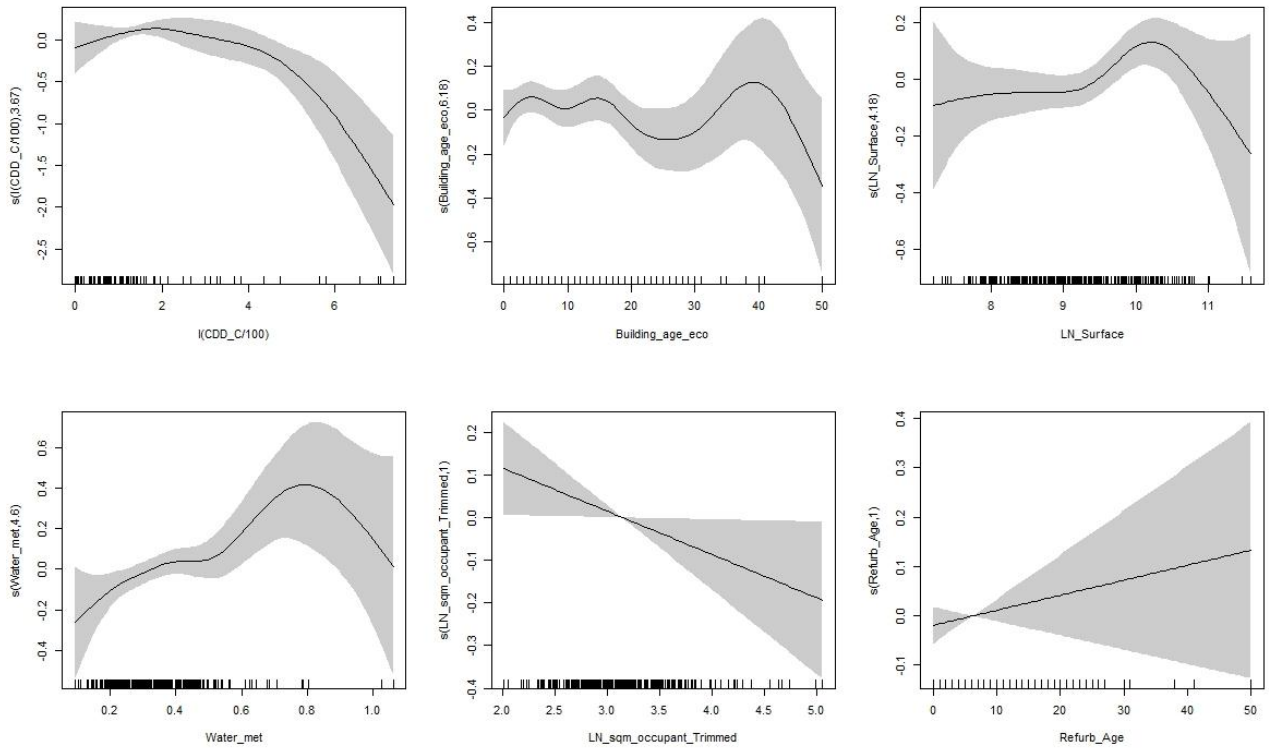
Significance: *** 1%, ** 5%, * 10%

The regression spline for *CDD* indicates only limited observations for the small data-set and only few observations with a higher annual number of *CDD*, thus indicating instability in the effect. However, the effect is significant for very low number of *CDD* (between values 0 and 2 on the x-axes), indicating slightly higher energy consumption when *CDD* are increasing due to higher air-condition loads.

The regression spline for building age shows high volatility in the effect. The highest energy consumption is estimated for office buildings around 15 years of age, whereas the most recent buildings consume slightly less energy, potentially with reference to higher energy efficiency. Older buildings with more than 20 years of age are attributed with significant lower energy consumption. The effect is significant up to the age of 30 years, followed by instability with limited number of observations. At first glance, this suggests that observations of building age lower than 15 years indicate higher consumption in regard to additional

appliances and equipment. Older buildings are assumed to provide less services and comfort, therefore associated with lower consumption, since we cannot distinguish between refurbished and non-refurbished observations in this spline.

**Figure 3: Regression splines of log-linear model on actual energy consumption as response variable from equation (3)**




The regression spline for the building size is supportive to our interpretation of the linear effect. Up to a building size of approx. 12,000 square metres, the effect is almost indifferent for smaller observations. For buildings of more than 12,000 square metres, a strong and significant increase in energy consumption is observable with further increase in the building area. The effect might result from high-rise office towers in the database with higher levels of energy consumption, due to large vertical distances. The results prove not for any potential economies of scale in energy consumption in reference to the building size.

Higher intrinsic water consumption assessment corresponds to significantly higher actual energy consumption up to a certain level, indicated in the spline. Energy and water efficiency are correlated to each other. This could reflect the stricter efficiency regulation and building codes, indeed affecting the physical building. Further explanation could be that for most recent constructed buildings, rated in the Green Rating audit, the modelled intrinsic water consumption is set low when new buildings in fact consume lower energy.

The regression spline for office area allocated per occupant proves for a linear relationship (*edf* = 1.0) in the effect, corresponding to the significant effect found before in the linear specification of the regression model. More space per occupant allocated is attributed with lower energy consumption per square metre. We found this corresponding to the results of Kahn et al. (2014), who observed that an increase in the occupancy rate increases the electricity consumption of office buildings.

The regression spline for the interaction effect between building age refurbishment illustrates also a linear relationship with highest energy efficiency for the most recent refurbished buildings. Since this effect is to be interpreted as the additional energy consumption of refurbished observations in relation to the building age and besides the main effect of refurbishment, introduced with a dummy for full / major / general refurbishment, it turns out that with higher age (as a proxy for the time when full / major / general refurbishment was carried out), the additional energy consumption of the buildings is significantly increasing. In general, younger observations have significantly lower energy consumption than older buildings. For the younger peers, although assumed with additional services and equipment introduced upon refurbishment, it seems that they are more energy efficient than their older peers. This indicates a less energy-efficient refurbishment for older buildings carried out in the past, but with limited observations for a building age of more than 25 years and a rather board confidence cone. The repression spline suggests that refurbishment from the past (more than 10 years ago) was less dedicated to energy-efficient appliances and equipment, compared to most recent refurbishment efforts, all else equal. This result for lower additional energy consumption in relation to refurbishment and lower building age might reflect the influence of stricter building codes and more efficient design of office buildings, implemented over the last years.

To sum it up, the introduced *GAM* visualized with regression splines provides a more precise understanding of the effects for the tested portfolio from *GRA*.

## 2.6 Conclusion

The objective of this study was to determine the influence of physical building characteristics and of occupant attributes on the actual energy consumption of European office buildings contained in the database of the *GRA* with 288 observations. Furthermore, we analysed the role of refurbishment on energy consumption with the effects of an extensive refurbishment and refurbishment solely dedicated to energy efficiency and thermal building characteristics.

Besides the application of a regression model to estimate linear effects on actual energy consumption, the study introduces *GAM* and visualisation with smoothed curves in regression splines to control for expected nonlinearity in the effects of the regression.

The assumed impact of occupant attributes on energy savings was shown to apply for the distinction between single and multi-tenant buildings. Single tenant buildings have a higher actual energy consumption between 12 and 15% compared to multi-tenant buildings (depending on the applied model specification). The result indicates a less energy-efficient behaviour of single tenants responsible for one building as a whole.

The electric production of heating is estimated with significantly lower energy consumption per square metre of up to more than 20%. It could be the case that highly efficient technology for heating production is employed in office buildings which are overall attributed with a relatively high energy efficiency standard in the tested portfolio. Given the fact that the data-set has almost 200 observations located in France, of which more than 50% are equipped with electric heating production, this result might point to higher energy efficiency employed with electric production of heating, compared to omitted reference production types gas, fuel or a district heating network.

In the regression model with linear effects, significantly lower energy consumption is found for observations with more than 20 years of building age. The introduced specification with *GAM* allows a

more detailed interpretation of the effect with highest energy consumption for office buildings around 15 years of age. For more recent buildings, slightly lower energy consumption is observed, potentially indicating higher energy efficiency in regard to increased energy efficiency regulation in *EU*-member states over the past decade. However, these attempts do not seem to reveal a significant impact on the conservation of actual energy consumption, based on the observations of the *GRA* portfolio. Corresponding to the principle of additionality with more and new appliances and equipment in more recent office buildings, we found older buildings of more than 20 years up to 30 years of age associated with significant lower energy consumption. These buildings are assumed to provide less services and comfort, which seems to correspond to observations for US office buildings.

Besides the linear coefficients introduced to control for the classified building size, the regression spline is of much higher explanatory quality for the nonlinear effect. The energy consumption of office buildings up to size of approx. 12,000 square metres is identified with a zone of almost indifference, followed by significantly higher energy consumption with increasing building area. For observations around 30,000 square metres, the highest consumption is observable. We find this in line with results of Kahn et al. (2014) when arguing that higher consumption is achieved in office towers, to bridge large vertical distances than in more compact building structures – rejecting the assumption of economies of scale.

Regarding actual energy consumption, the high explanatory power of the intrinsic water consumption is an interesting result of the study. Since this intrinsic measure is modelled according to the technical building characteristics and is independent of actual building use, it provides a potential indicator for high energy efficiency. This could reflect stricter efficiency regulation and building codes, indeed affecting the physical building. From an investor's point of view, this underlines the importance of measuring further sustainable characteristics besides energy, when high water efficiency is identified to be an indicator of high energy efficiency from the *GRA* data-set.

The linear coefficient but even more the regression spline for the allocated office space per occupant proves for a significant relationship with lower energy consumption per square metre when more office space is allocated per occupant. The regression spline indicates a linear effect in the relationship. Besides the effect of higher vacancy in office buildings associated with lower consumption, this also suggests that more occupants on the same office area will increase energy use per square metre from applied equipment and plug loads (especially if the building is heated or cooled centrally).

Since the role of refurbishment is of major research interest in this study, the attributes for extensive refurbishment (full / major / general) proved for significantly higher energy consumption of 15% as a linear covariate effect, compared to buildings without any refurbishment measures. For the tested portfolio, we find the principle of additionality being associated with refurbishment that provides increased or additional services and comfort from appliances with additional or new equipment. Since the extensive refurbishment was dedicated to improve the overall quality of the buildings, the measures might include those of higher energy efficiency, but from the additional appliances and equipment, the overall energy consumption is higher compared to peers without refurbishment.

Refurbishment dedicated solely to energy efficiency and thermo-physical building characteristics remains insignificant in the second specification of the regression model. The coefficient is estimated with a negative value, corresponding to potential energy savings, but insignificant, most probably because of only 23 observations. Since the effect is not estimated with a positive coefficient, there is no indication of a potential rebound effect inherent in the portfolio at all.

Finally, the regression spline for the interaction effect between building age and refurbishment, to be interpreted as additional energy consumption of refurbished buildings, shows the highest energy efficiency for the most recent refurbished buildings, although assumed with additional services and equipment introduced upon refurbishment. While the regression spline indicates a less energy efficient refurbishment for older buildings carried out in the past, this suggests that refurbishment from the past was less dedicated to energy- efficient appliances and equipment, compared to most recent refurbishment efforts, all else equal. This result for lower additional energy consumption in relation to refurbishment with lower building age might reflect stricter building codes and more efficient design of office buildings, implemented over the last years.

In the discussion and conclusion of our study results, we repeatedly referred to the small database from *GRA* of only 288 office buildings in 14 European countries as a major drawback.

As of now, there is no extensive research framework for energy consumption of commercial, especially office buildings, based on empirical evidence from Europe. This study aims to contribute to a better understanding of energy consumption in office buildings, in particular, when it comes to the role of refurbishment with differing intensity. However, there are limitations remaining from the small data-set, which lacks higher explanatory power. A reason for the very moderate fit of the regression models, indicated in the relatively low $R^2$ values and rather weak significance of the attributes might be due to omitted variables in our models or defaults in the used data.

Since the used data-set contains the actual metered energy consumption and the intrinsic assessment only assessed once for issuing the Green Rating audit, the results of the study should be interpreted with precaution. According to the limited nature of the data-set, we cannot exclude the possibility that any potential bias due to omitted variables is included in the outcomes. Besides the rather small number of observations from different years between 2008 and 2012, we do not have further information in regard to the building quality and maintenance or the equipment and appliances, differing among the observations. Furthermore, the missing information in regard to the sub-category 'other' in the data-set is a limiting factor that could reduce explanatory power and introduce omitted variable bias. Moreover, we do not know the vacancy rate in the buildings or unemployment rates for the relevant time frame. Also, the business industries of the office occupants might be associated with potential differences in energy consumption, compared to each other. Another limitation is that we do not know and control for the relevant energy cost in the buildings. However, as an advantage, the used data-set contains detailed information to differentiate among the intensity of refurbishment measures, which is a precondition to investigate the role of refurbishment on energy consumption.

We tried to mitigate the problem of nonlinearity in the effects of the estimated regression coefficients with the introduction of *GAM*. The obtained regression splines provide a more precise understanding of the effects for the observed portfolio. However, a dynamic analysis of a panel with more observations over time – controlling the effects before and after refurbishment – is recommended.

Beyond the recommendation to consolidate our research results on a more extensive database, we believe that fostering the awareness of actual energy consumption in the direction of potential savings by occupants is furthermore an issue. Apart from technological progress and a more extensive and stricter energy efficiency regulation, the design of an effective (incentive) mechanism to shift office occupant behaviour towards energy conservation might achieve higher energy savings in practice. How this

mechanism could be designed is a subject for further research, especially in the field of behavioural real estate research.

## 2.7 References


Association of European Renewable Energy Research Centres (Ed., 2016). Online: http://www.eurec.be/en/ (accessed: 15.11.2015).

Baum, A., Turner, N. (2004). Retention Rates, Reinvestment and Depreciation in European Office Markets, *Journal of Property Investment and Finance,* 22(3), 214-235.

Baum, A., McElhinney, A. (1997). The Causes and Effects of Depreciation in Office Buildings: A Ten Year Update. Report for working paper of the Department of Land Management and Development, University of Reading.

Bloom, N., Genakos, C., Martin, R., Sadun, R. (2011). Modern Management: Good for the Environment, or Just Hot Air?, *Economic Journal,* 120(544), 551-572.

Bouwfonds (Ed., 2013). Corporate Social Responsibility Policy. Bouwfonds Investment Management, online: http://www.bouwfondsim.com/wp-content/uploads/2014/10/BIM_CSR-Policy_14112013.pdf (accessed: 15.11.2015).

Brunauer, W. A., Feilmayr, W., Wagner, K. (2012). *A new residential property price index for Austria*. Statistiken – Daten und Analysen Q3/2012, Austrian National Bank, Vienna, 90-102.

Brunauer, W. A., Lang, S., Wechselberger, P., Bienert, S. (2010). Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna, *Journal of Real Estate Finance and Economics,* 41(4), 390-411.

Cajias, M., Bienert, S. (2011). Does Sustainability Pay Off for European Listed Real Estate Companies? The Dynamics between Risk and Provision of Responsible Information, *Journal of Sustainable Real Estate*, 3(1), 211-231.

Cajias, M., Fuerst, F., McAllister, P., Nanda, A., (2011). Is ESG commitment linked to investment performance in the real estate sector? Working Papers in Real Estate and Planning, 08/11. Working Paper. University of Reading.

Chegut, A., Eichholtz, P., Kok, N. (2015). The Price of Innovation: An Analysis of the Marginal Cost of Green Buildings. Working paper, available at: http://www.corporate-engagement.com/files/publication/CEK_Cost_102315.pdf (accessed: 22.11.2015).

Fahrmeir, L., Kneib, T., Lang, S. (2007). *Regression. Modelle, Methoden und Anwendungen*. Springer, Berlin.

Fraunhofer IBP (Ed., 2009). *Studie zur Energieeffizienz innovativer Gebäude-, Beleuchtungs- und Raumkonzepte (EnEff-Studie)*. Fraunhofer-Institut für Bauphysik (IBP), Holzkirchen.

Fraunhofer ISE (Ed., 2015). Fraunhofer-Institut für Solare Energiesysteme ISE, online: https://www.ise.fraunhofer.de/en/about-us (accessed: 01.04.2016).

Gabe, J. (2014). Successful greenhouse gas mitigation in existing Australian office buildings, *Building Research & Information,* 44(2), 160-174.

Gillingham, K. M., Rapson, D. S., Wagner, G. (2015). The Rebound Effect and Energy Efficiency Policy, *Review of Environmental Economics and Policy,* 10 (1), 68-88.

Gillingham, K. M., Kotchen, M. J., Rapson, D. S., Wagner, G. (2013). The Rebound Effect is overplayed, *Nature, 493*, 475-476.

Greening, L, Greene, D. L., Difiglio, C. (2000). Energy Efficiency and Consumption – the Rebound Effect – a Survey, *Energy Policy, 28(6-7)*, 389-401.

Guerra Santin, O., Itard, L., Visscher, H. (2009). The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock, *Energy and Buildings,* 41(11), 1223-1232.

Halvorsen, R., Palmquist, R. (1980). The Interpretation of Dummy Variables in semilogarithmic equations, *The American Economic Review,* 70(3), 474-475.

Hardy, M. A. (1993). *Regression with Dummy Variables*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-093. Sage, Newbury Park, CA.

Herring, H., Roy, R. (2007). Technological innovation, energy efficient design and the rebound effect, *Technovation,* 27(4), 194–203.

Jenkins, D., Liu, Y., Peacock, A. D. (2008). Climatic and internal factors affecting future UK office heating and cooling energy consumptions, *Energy and Buildings*, 40(5), 874–881.

Kahn, M. E., Kok, N., Quigley, J. M. (2014). Carbon Emissions from the Commercial Building Sector: The Role of Climate, Quality, and Incentives, *Journal of Public Economics,* 113, 1-12.

Kavulya, G., Becerik-Gerber, B. (2012). Understanding the influence of occupant behavior on energy consumption patterns in commercial buildings. International Conference on Computing in Civil Engineering, (Proceedings), 569–576.

Kerscher, A. N., Schäfers, W. (2015). Corporate social responsibility and the market valuation of listed real estate investment companies, *German Journal of Real Estate Research*, 1(2), 117-143.

Knittel, C. R. (2012). Automobiles on Steroids: Product Attribute Trade-Offs and Technological Progress in the Automobile Sector, *American Economic Review,* 101(7), 3368-3399.

Kok, N., Miller, N. G., Morris, P. (2012). The Economics of Green Retrofits, *Journal of Sustainable Real Estate,* 4(1), 4-22.

Lombard, L. P., Ortiz, J., Pout, C. (2008). A review on buildings energy consumption information, *Energy and Buildings,* 40(3), 394-398.

Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review, in: O'Sullivan, T., Gibb, K. (Eds.), Housing Economics and Public Policy, Essays in Honor of Duncan Maclennan, 67-89, Blackwell, Oxford.

Mason, C., Quigley, J. M. (1996). Non-parametric Hedonic Housing Prices, *Housing Studies,* 11(3), 373-385.

New Buildings Institute (Ed., 2008). Energy Performance of LEED for New Construction Buildings. Prepared for: U.S. Green Building Council. New Buildings Institute, Vancouver, WA.

Newsham, G. R., Mancini, S., Birt, B. J. (2009). Do LEED-certified buildings save energy? Yes, but…, *Energy and Buildings,* 41(8), 897-905.

Pace, R. K. (1998). Appraisal Using Generalized Additive Models, *Journal of Real Estate Research,* 15(1), 77-99.

Parmeter, C. F., Henderson, D. J., Kumbhakar, S. C. (2007). Nonparametric estimation of a hedonic price function, *Journal of Applied Econometrics,* 22(3), 695-699.

Chartered Institution of Building Services Engineers (Ed., 1997). Post-occupancy Review of Buildings and their Engineering. PROBE 9: Energy and Engineering PROBE Technical Review. online: http://www.cibse.org/Knowledge/Building-Services-Case-Studies/PROBE-Post-Occupancy-Studies (accessed: 28.02.2016).

Scofield, J. H. (2009). Do LEED-certified buildings save energy? Not really…, *Energy and Buildings,* 41(12), 1386-1390.

Surmann, M., Brunauer, W., Bienert, S. (2016). The energy efficiency of corporate real estate assets: The role of energy management for corporate environmental performance. *Journal of Corporate Real Estate*, Special Issue from the 22nd European Real Estate Society (ERES) annual conference, Istanbul, Turkey.

Torcellini, P. A., Deru, M., B. Griffith, B., N. Long, N., S. Pless, S., R. Judkoff, R. (2004). *Lessons learned from the field evaluation of six high-performance buildings*. ACEEE summer study on energy efficiency of buildings, Conference Paper, National Renewable Energy Laboratory, Golden, Colorado, USA.

Weather underground (Ed., 2015). Online: http://www.wunderground.com (accessed: 22.07.2015).

Wood, S. (2006). *An Introduction to Generalized Additive Models with R*. Taylor & Francis Group, Boca Raton, FL.

# 3 Assessment of climatic risks for real estate

**Jens Hirsch** **Thomas Braun** **Sven Bienert**

**Structured Abstract:**

**Purpose**: The paper investigates the functionality and main results of the *ImmoRisk* tool. The aim of the project of the *Federal Ministry for Transport, Building and Urban Development* (*BMVBS*), in corporation with the *Federal Institute for Research on Building, Urban Affairs and Spatial Development* (*BBSR*), was to develop a user-friendly tool that provides a sound basis with respect to the risk situation caused by extreme weather events.

**Design / methodology / approach:** The tool calculates the annual expected losses (AEL) for different types of extreme weather hazard and the damage rate as the proportion of AEL on building value, based on a trinomial approach: natural hazard, vulnerability and the value of the property.

**Findings:** The paper provides property-specific risk profiles of both the present and future risk situation caused by various extreme weather events.

**Research Limitations/ implications:** The approach described in the paper can serve as a model for the realization of subsequent tools in further countries bound with other climatic risks.

**Practical implications:** The real estate industry is affected by a significant rise in monetary damages caused by extreme weather events. Accordingly, the approach is suitable for implementation in the companies' real estate risk management systems.

**Social Implications:** The tool offers homeowners a profound basis for investment decisions with regard to adaptation measures.

**Originality/ value:** The approach pioneers fourfold, (1) by meeting the needs of the housing and real estate industry based on a trinomial approach, (2) by using a property-specific bottom-up approach, (3) by offering both a comprehensive risk assessment of the hazards storms, flood and hailstorm, and (4) by providing results with respect to the future climatic risk situation.

**Keywords:** Risk Analysis, Extreme Weather Events, Climate Change, Property Damage

**Article Classification:** Conceptual paper

**Acknowledgments:** The authors are grateful to Ute Birk and Tobias Held from the *Federal Institute for Research on Building, Urban Affairs and Spatial Development* (*BBSR*) for initiating and supporting the *ImmoRisk* project, Guido Halbig, Bruno Merz, Marita Roos and Martin Vaché and the members of the steering committee for their scientific support, the *German Weather Service* (*DWD*), the *German Research Centre for Geosciences* (*GFZ*), the *Karlsruhe Institute of Technology* (*KIT*) and companies of the insurance industry for providing the crucial data set. They also express their appreciation to Marcelo Cajias and Peter Geiger for their valuable research assistance, Brian Bloch for his comprehensive editing of the manuscript, several conference audiences for helpful advices. The authors received funding from the *German Society of Property Researchers* (*gif*) and the *Vielberth Foundation*. They express their gratitude for this financial support.

## 3.1 Introduction

According to reinsurance companies, both the quantity and intensity of extreme metrological events have increased significantly over the past three decades, e.g., the global number of meteorological and hydrological events like storms or flooding has increased from 1980 to 2013 by more than 250 percent (Munich Re, 2014). Focusing on Germany, the damage caused by meteorological (storm and hail), hydrological (flood and mass movement) and climatologic events (extreme temperatures, droughts and forest fire) between 1980 and 2011 is estimated at €70 billion, i.e., €2.2 billion yearly, less than 50 percent being covered by insurance. Climatologists confirm this trend and forecast a further intensification for the upcoming decades, which will not least be reflected in higher insurance costs to cover the damage.

The emerging character of climate change demands both adaptation and mitigation strategies. While mitigation measures in terms of energy-efficient buildings are omnipresent, adaptation measures in the construction sector have not reached a similar level. Although the housing and real estate sector is engaged in climate change, natural hazards are underrepresented in the industry's risk analyses. "*The availability of information, e.g. about monitoring systems and guidelines, will be an important basis for adaptation and climate protection measures in the building sector*" (German Federal Government, 2008, p. 16). However, there is a lack of small-scale data sets applicable to the housing and real estate industry, which focus the frequency and intensity of extreme weather events like storms, heavy precipitation or hailstorms. Particularly small and medium-sized enterprises (*SME*) or private house-owners experience difficulties. As their properties and especially their sites are fixed, they have only little potential for diversification. But even institutional investors need reliable information about increasing natural hazards and their impact on assets. The actors in the real estate sector require quantitative and particularly future-oriented data, which they can integrate into their risk information systems. Existing meteorological parameters do not enable the estimation of expected damage on their own and have to be customized to the specific requirements of the real estate industry. Otherwise, cost-benefit analyses of adaptation measures are less reliable, which might lead to capital misallocation. Therefore, the German Federal Ministry for Transport, Building and Urban Development (BMVBS), in corporation with the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR), launched the project *ImmoRisk*. It has taken a first step in reducing the prevailing information deficit by making property-specific risk analysis publicly available for the first time, and covering a multitude of locations in Germany.

A trinomial approach was selected, in order to develop the first tool that quantifies climate risks caused by storms, flooding and hailstorms for individual properties within different timeframes. It covers the components natural hazard, property vulnerability and monetary value. Hazard data were collected for 15 sites from the *German Weather Service* (*DWD*), the *German Research Centre for Geosciences* (*GFZ*) and the *Karlsruhe Institute of Technology* (*KIT*), which ensured state of the art data from regional climate models. Vulnerability is modeled with damage functions, based on existing literature and from the insurance industry. Finally, the data set were completed by customized standard construction cost data provided by the Federal Ministry of Building. After entering user-specific property input, the tool calculates the annual expected loss (*AEL*) in Euros and the damage rate for storms, flooding and hailstorms for individual properties. Additionally, the output comprises information on the hazard of the location with respect to forest fires, heat, heavy precipitation,

lightning, as well as excess voltage. For validation purposes, the results are compared with the level of corresponding insurance policies.

The remainder of the paper is organized as follows. Section 3.2 presents the existing tools for the quantification of future property climate risks. Section 3.3 explains the general conceptualization of the *ImmoRisk* tool and its specific understanding of risk provided by a trinomial approach containing natural hazards, vulnerability and property value. Section 3.4 describes the technical implementation of the web-based platform *ImmoRisk* with regard to the data used and technological solutions. Section 3.5 comprises the functional implementation of the *ImmoRisk* tool from the user perspective including an overview of step-by-step risk assessment, results and other functions. Finally, Sections 3.6 and 3.7 present the results of the risk assessments conducted with the *ImmoRisk* tool and outline future challenges with regard to adaptation strategies in the real estate industry.

The *ImmoRisk* project aimed to develop a real estate specific tool for identifying risks caused by natural hazards. Players in the real estate industry were intended to obtain practice-relevant information applicable to existing risk information systems. For this purpose, 15 pilot locations were selected. These locations are distributed throughout Germany and cover different climatic conditions in urban as well as rural regions. The plot size of the examined locations varies between 200m× 200m and 700m×700m.

## 3.2 Background

### 3.2.1 Need for climatic risk assessment

Investors require information with respect to the natural risk of a specific property, ideally considering risk-relevant constructional features. This applies both, to the acquisition of properties as well as to investments in individual adaptation measures. Hence, highly spatially aggregated data does not offer sufficient action-relevant information. Additionally, most existing platforms have in common that they cover only the current climate, i.e., they do not provide information on the adjusted risk situation caused by climate change. For purposes of real estate risk management, it is essential that all natural hazards are covered, which have significant influence on the risk situation of a property. A singular consideration of separate natural hazards therefore cannot be sufficient. Accordingly, a respective tool for identifying natural risks in the German-speaking area must at least cover damage caused by wind storms, flood and hailstorms. The hail season 2013 caused damage of approximately €3.1 billion to the property insurance companies. This considerably exceeds the damage resulting from wind storm and flood (GDV, 2014). As just parts of the damage were covered, the actual economic costs are assumed to be significantly higher.

In summary, a risk information system of natural hazards should comply with the following conditions in order to meet the needs of the housing and real estate industry:

(1) alignment of quantitative, monetary data covering expected damages according to the requirements of the real estate industry,
(2) information at the property level,
(3) consideration of the vulnerability of an individual building or even just parts of it,
(4) consideration of all natural hazards in the examined area,
(5) consideration of climate change and
(6) public availability.

The last-mentioned aspect of public availability accounts for the fact that (re)insurance companies and some providers of specialized business solutions developed own internal software for risk identification, which however is not made available for public. This lacking market transparency leads to a privilege for individuals with data access, in this case, of institutional investors. A professional real estate company may use the services of special providers for optimizing their own portfolios, but neither ordinary homeowners nor smaller communities are typically customers of these services. In order to ensure a maximum of market transparency and to achieve an optimal result of capital allocation, a natural risk information system should be available for public use without limitations.

### 3.2.2 Existing tools for risk assessment

Existing tools for estimating natural hazards do not cover sufficiently the specific requirements of the real estate industry. Solely in the German-speaking area, numerous platforms such as *ZUERS Public* (www.zuers-public.de/zuerspublic/) or *eHORA – Natural Hazard Overview & Risk Assessment Austria* (www.hora.gv.at/) exist that provide information with regard to the statistical frequency of certain natural phenomena and their spatial distribution. However, retrieved maps usually focus on displaying information about the relative hazard to regions on an ordinal scale or the frequency of events of a given level of intensity. Hence, this just sketches an outline, as action-relevant data with regard to the property-specific damage risk are lacking. International platforms like *RiskMeter Online* (*Core Logic*) or *CatNet* (*Swiss Re*) also do not provide monetary information at property level.

Furthermore, a variety of platforms exist, which offer information on current threats or short-term forecasts. Their information is relevant for rescue forces or homeowners. However, these platforms do not fit the requirements of real estate risk management systems.

Only few platforms such as the *CEDIM Risk Explorer* (http://cedim.gfz-potsdam.de/riskexplorer/) exceed the pure observation of natural events and connect them to the resulting damage risk. However, these approaches are dedicated to the point of view of insurance companies. Therefore, these platforms operate on a high level of spatial aggregation. The *CEDIM Risk Explorer* offers information on expected damage by winter storms (50-, 100- and 500-year events), earthquakes (475-year event) and flooding (200- to 500-year event) on a zip code level. In contrast, the housing and real estate industry requires information at the property level.

## 3.3 Conceptualization of the ImmoRisk tool

### 3.3.1 Trinomial risk approach

The main challenges of developing a property-specific tool for identifying natural risks of the real estate sector were the implementation of a user-friendly technical solution and the functional combination of climate data and information about property vulnerability from a variety of sources. In addition, the system should allow users to record multiple properties and ensure the confidentiality of user data.

From a management point of view, risk expresses an action situation with an uncertain result, which can imply both negative and positive deviations from the expected value. For example, while credit risks can lead to unexpected profits, i.e., upside risk, downside risk is more apparent in the context of extreme meteorological events.

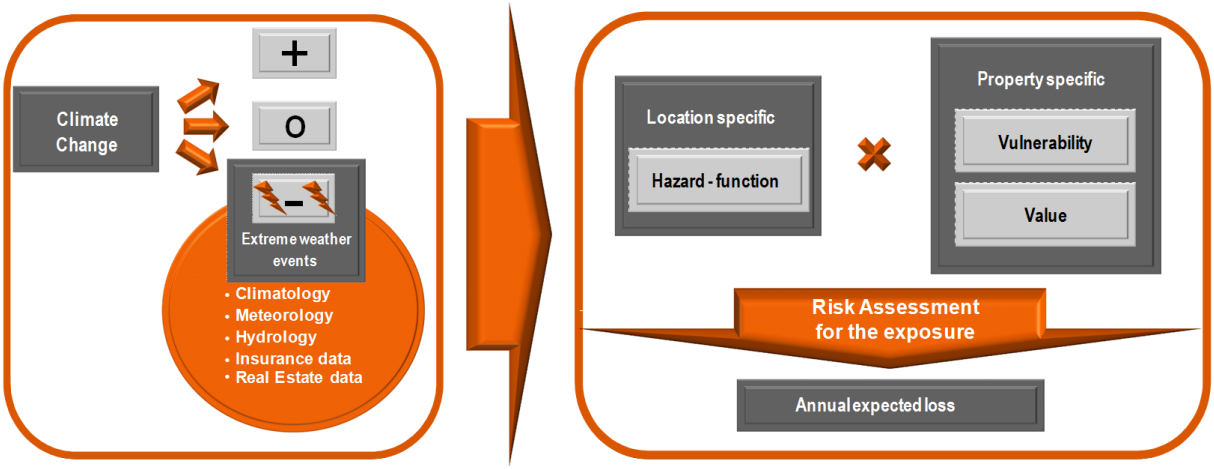
For the purposes of the *ImmoRisk* project, risk as a term was defined as the probability of occurrence of specific damaging events. In order to achieve a maximum of action relevance for real estate practice, risk is defined as the expected damage or loss on an annual basis.

Meteorological key figures, which describe the probability of occurrence of natural hazards, do not directly enable the determination of property risk. The intensity of a natural event, such as the wind speed of a storm or the water depth of a flood, does not reveal anything about the expected damage, which depends highly on the characteristics of a specific property. The probability of occurrence of a natural event with a particular intensity at a certain location is defined by the term 'hazard'. Potential methods to model the hazard are described in Section 3.3.2. The sensitivity of a property with specific physical characteristics is defined as vulnerability. Mathematically, vulnerability is expressed by so-called damage functions, which state a functional relationship between the intensity of a natural event and the resulting damage. The risk associated with a property can therefore be described as a combination of a location-specific hazard and a property-specific vulnerability. Damage functions often do not refer to absolute, monetary values, but rather yield damage ratios. Depending on the intensity of the natural event, the damage accounts for a specific percentage of property value. The *ImmoRisk* tool therefore comprises, besides the parameters of hazard and vulnerability, also a valuation approach which is based on construction costs (see Figure 1):

$$Risk = Hazard \circ Vulnerability \circ Value \qquad (1)$$

In the following sections, the details for determining and modeling hazard, vulnerability and value are further outlined. Subsequently, we explain, how these three elements can be transformed into property risks.

**Figure 1: General approach for the assessment of real estate risks**




*Source: Bienert et al. (2013)*

## 3.3.2 Hazard

### 3.3.2.1 Hazard statistics

In the present case, hazard means the probability that natural events of a certain intensity take place. This probability is described by the so-called hazard function. Only by using the respective damage functions, statements with regard to the risk - being the probability of monetary damage -

can be made. Thus, for quantitative risk modeling, initially the hazard must be considered. The frequency of natural events that exceed a certain level of intensity is related to the so-called recurrence interval or return period.

The hazard function is determined by adapting suitable distribution functions to empirical historical data or data generated by climate models. The historical data includes, for example flood marks on buildings, metered values from meteorological stations or, very recently, remote sensing data provided by satellites or radar stations. In this context, a certain group of parametric statistical distributions is deployed, which originate in so-called extreme value statistics. These distributions reflect the statistical behavior of very rare events correctly. Other conventional distributions, like normal or exponential distribution, would underestimate the frequency of these rare events, although they cause tremendous damage.

Methods in extreme value statistics, which are employed in the frequency analysis of extreme meteorological events, include the *Generalized Extreme Value Distribution* (*GEV*), as well as the *Peaks Over Threshold Method* (*POT*), which leads to a *generalized Pareto distribution*. For a particular case, the decision to use one of these methods, must be linked to the best model fit to the empirical data. *GEV* and *POT* are parametric distributions defined by the three parameters of shape, location and scale. The estimation of such parameters is usually based on the methods of moments or maximum likelihood. The application of extreme value statistics is based on the extremal types theorem, which states that extreme values of an independent and identically distributed random variable converge asymptotically toward the *GEV* distribution.

*GEV* is based on the assumption of independent and identically distributed intensities of respective natural events. In order to determine the frequency of extreme meteorological events using *GEV*, the sample period is divided into identical periods, e.g., years, and only the strongest events in each period are used for estimating the parameters. All other events remain unconsidered. According to the *Fisher Tippett Gnedenko theorem*, the *GEV* is a generalized formulation of the *Gumbel*, the *Fréchet* and the *Weibull distribution*. The probability density in *GEV* can be written as:[10]

$$f(x,\mu,\sigma,\xi) = \frac{1}{\sigma}\left(1+\frac{x-\mu}{\sigma}\right)^{-1-\frac{1}{\xi}} exp\left(-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{\frac{-1}{\xi}}\right) \qquad (2)$$

with $x$, intensity of an event; $\mu$, location parameter; $\sigma$, scale parameter; $\xi$, shape parameter.

The *Gumbel distribution*, which is specifically used for describing intense storm events, depicts the special case of $\xi = 0$, so that only the parameters of $\mu$ and $\sigma$ have to be estimated. The *Fréchet distribution* is obtained for $\xi > 0$ and the *Weibull distribution* for $\xi < 0$ (Coles, 2001).

The return level $x$, which is statistically exceeded once every $1/p$ years, can be calculated using the estimated parameters as follows (Coles, 2001)[11]:

[10] Markose, Alentorn, 2011
[11] Derived from the inverse of the GEV distribution function

$$x(p) = \begin{cases} \mu - \frac{\sigma}{\xi}\left(1 - \left(-log(1-p)\right)^{-\xi}\right), for\ \xi \neq 0 \\ \mu - \sigma log\left(-log(1-p)\right), \qquad for\ \xi = 0 \end{cases} \quad (3)$$

with $p$, exceedance probability.

The exceedance probability can be interpreted as probability of exceeding the return level within one year.

Contrary to *GEV*, the *POT* method does not only consider a certain number of maxima of each period, but rather all events that exceed a certain threshold $u$ in their intensity. The value of the exceedance ($y = x - u$) can be interpreted as an independent realization of a generalized *pareto distributed* random variable (Coles, 2001).

For determining the frequency of extreme meteorological events, the parameters of the generalized *pareto distribution* are therefore selected in such a way that they describe the statistic behavior of the events beyond the threshold at the best. When determining the threshold, two contrasting phenomena have to be balanced. An underestimated threshold could result in a violation of the model assumption of asymptotic properties and the application of extreme value statistics implies biased results. On the other hand, an overestimated threshold results in only few events exceeding the respective value which increases the variance of the parameter estimations. The return levels $x$ as a function of the exceedance probability p results as:

$$x(p) = u + \frac{\sigma}{\xi}\left(1 - \left(\frac{n}{M \cdot p}\right)^{-\xi}\right) \quad (4)$$

with *n*, total number of events; *M*, number of years (Kunz and Puskeiler, 2010).

#### 3.3.2.2 Climate Modeling

As properties are durable goods and climate is changing, risk management is supposed to exceed the mere extrapolation of historical data. In order to identify the hazards at a location for future periods, data from climate models are used and evaluated by means of extreme value statistical methods. Initially developed global climate models or *General Circulation Models* (*GCM*) allow an adequate description of future climate development by linking several modules covering the atmosphere, the oceans, the cryosphere and several additional elements of the climate system. *GCM* divide the atmosphere into three-dimensional volume elements, for which respective meteorological variables are calculated. However, the spatial resolution of these models that illustrate the atmosphere of the entire planet is too low for matching the hazard to properties, as extreme weather events occur with high-spatial variance. As regional climate models provide resolutions of less than 10km×10km, these are applied in the *ImmoRisk* project. Thereby, they model climate conditions for areas of continental or sub-continental extent and transfer the results of *GCM* using different methods to a grid of higher resolutions, i.e., 'downscaled'. As humankind was identified as an essential driver of the current climate change[12], uncertainty within the pathways of the future anthropogenic influence on the

[12] "More than half of the observed increase in global mean surface temperature (*GMST*) from 1951 to 2010 is very likely due to the observed anthropogenic increase in greenhouse gas (*GHG*) concentrations." (IPCC, 2013, p. 885). The term 'extremely likely' is used by the *IPCC* in order to indicate an assessed likelihood > 95%.

climate are taken into account by scenarios. These scenarios describe different assumptions, each with regard to the development of greenhouse gas emissions, determined by the economic and demographic development, as well as the adoption of technological innovations. Climate modeling that considers a specific scenario is referred to as climate projection.

Two different groups of *RCM* can be distinguished, *dynamic RCM* (*dRCM*) and *statistic RCM* (*sRCM*). Analogous to *GCM*, *dRCM* aim to depict processes in the atmosphere by means of thermodynamic laws of conservation. The possibility to consider parameterized convection processes constitutes an advantage of *dRCM*, but the extremely high-computational costs of this method limit the number of model runs.

On the other hand, *sRCM* forego an exact thermodynamic description of the atmospheric processes and revert to measurement data of a maximum number of meteorological stations. The basic approach of *sRCM* for projecting future climate is creating a link between meteorological parameters of past periods, for example temperatures or frequencies of typed weather situations, and observational data of the meteorological stations. The detected relationship between observed weather and certain variables is used to derive future climate for each station by using the projections of these variables from GCM simulations. For this approach, a preferably dense network of measuring stations that deliver a sufficiently long series of observational data are required. Hence, *sRCM* deliver values for a variety of atmospheric parameters in a high-chronological resolution for each measuring station. Applying spatial interpolation techniques, the results can be transferred to a discrete spatial grid, comparably to *dRCM*. In contrast, the advantages of *sRCM* are a lower dependence of *GCM* results, as well as significantly lower computational costs per run, which enables more precise conclusions with respect to the variance of the projections (Sauer, 2010).

For the territory of Germany, two *dRCM* and two *sRCM* in high-spatial resolution are available, i.e., *CCLM*, *REMO* as *dRCM* and *WETTREG*, *STARS* as *sRCM*.

### 3.3.3 Vulnerability

The linkage of location-specific hazard and property-specific vulnerability is realized by means of damage functions. They provide a functional connection between intensity of a natural event and the occurrence of property damage. Scientific publications and data from insurers serve as sources for damage functions. Basically, there are two possibilities for deriving damage functions.

Empirical approaches are based on historical damage events, for which all damage is recorded afterwards and confronted with local intensities of the natural hazard, for example wind velocity or inundation depth (Merz et al., 2004). Empirical damage functions are available for the three most important climate risks in Germany, which are flood, storm and hail. The information about hazard intensities is derived from meteorological or hydrological measuring stations, sometimes in combination with modeled data for example about the exact topography of inundation. The damage data are based either on systematic interviews of affected people (Thieken et al., 2008) or on evaluations of insurance data (André et al., 2013). In most cases, the damage is indicated as a proportion of the insurance value of the property. The advantage of these relative damage functions lies in the fact that no constant adjustment to current price indices is necessary (Friedland, 2009, p. 25). Existing studies have resulted in a broad range of diverging damage functions for the same

natural hazard. Jongman et al. (2012) conducted a sensitivity analysis of seven flood damage functions and found a great variance especially concerning low levels of flooding.

Contrary to empirical damage functions, deterministic damage functions, partially also called synthetic, are based on an engineering-oriented approach. In this context, specific processes that damage building components are identified. For different building types, intensities are connected to different partial damage processes, based on both empirical experience and theoretical reflection (Naumann et al., 2009). Experts finally estimate the clean-up costs according to systematic procedures. Deterministic damage functions are particularly derived for storm damages in the *USA* (Unanwa et al., 2000), which are not automatically transferable to other regions, as both the construction and the costs of damage removal differ widely. In the case of Germany, there are some approaches to derive deterministic damage functions with regard to flood, as well (Naumann et al., 2009).

### 3.3.4 Value

In order to transform the relative damage in an absolute number, the trinomial structure of the risk assessment is completed by the value of the property. Existing large-scale risk assessments apply top-down calculations, in order to estimate exposed assets (see Thieken et al., 2006; Kleist et al., 2006) and derive corresponding risks (Heneka et al., 2006). Since climatic research institutes offer new high-resolution hazard data and the real estate industry has a strong need for property-specific risk assessments, the development and application of property-specific bottom-up approaches are reasonable.

As the cost approach is based on construction costs, it is the most appropriate one among the existing real estate valuation methods with respect to the assessment of damages caused by climatic risks (Heneka et al., 2006). It addresses neither a market-based willingness to pay nor reproduction costs in terms of a completely identical building, but rather replacement costs (Kleiber et al., 2010; RICS, 2009). Since the replacement of damages by age-related depreciated building components is almost impossible, the replacement term describes the components in mint condition prior to damage (Naumann et al., 2009). Besides omitting age-related depreciation, the used approach opposes the conventional cost approach in additional aspects. For example, the land value is not taken into account, as the parcel is rarely affected by damages caused by extreme weather events. Furthermore, damages with respect to furniture and equipment are also excluded from consideration like outdoor installations, since they are not covered by storm and tempest insurances, which form the basis of the used damage functions.

The cost approach assumes cost similarity of usage types, which is a prerequisite for calculating averaged costs based on empirical values. In Germany, standard construction costs are provided by the Federal Ministry of Building and are classified by region, type of use, type of construction, period of construction and quality of the property. In the *ImmoRisk* tool the corresponding categories are selected by user input with regard to structural features. As the standard construction costs are a static magnitude, they must ultimately be adjusted according to the construction cost index based on the reference date. The result is defined as replacement value based on usual construction costs per unit of gross floor area at the relevant valuation date.

### 3.3.5 Risk

The risk of a property at a certain location is a result of the factors hazard, vulnerability and value. For real estate risk management purposes, data with regard to expected damages e.g., by a one-in-a-century flood, are not sufficient. Additionally, information on different recurrence intervals has to be integrated resulting in an *AEL*. In this context, damage resulting from events of a certain recurrence interval influences the *AEL* on a weighted basis considering the exceedance probability. For example, observing a natural hazard described by the *Gumbel distribution* and an arbitrary damage function $D(p)$, the $AEL$ for a property with a value $V$ can be described as follows:

$$AEL = V \int_0^1 D\left(\mu - \sigma log\left(-log(1-p)\right)\right) dp \qquad (5)$$

## 3.4 Technical implementation of the ImmoRisk tool

### 3.4.1 Hazard and vulnerability data description

#### 3.4.1.1 Storm

In order to model storm hazards, the *ImmoRisk* tool draws on data of the *dRCM COSMO-CLM*. This reflects the maximum wind gust velocities on a daily basis, which are analyzed by extreme value statistical methods using *Gumbel distribution*. Current hazard is calculated from data of the period between 1971 and 2000, future hazard comprises the years 2021-2050. For both timeframes four model runs were performed based on the *IPCC* scenario *A1B*. Three of four runs refer to *GCMECHAM5-OM*, the remaining run is based on the *GCM CCCma3*. The parameters for the *Gumbel distribution* are available with a spatial resolution of approximately 7km×7km and are provided by the *Institute for Meteorology and Climate Research* at *KIT*.

The vulnerability modeling approach is based on data provided by *Deutsche Rueck* reinsurance company. In total, five damage functions were derived from the data, differing in the degree of vulnerability. The selection of the appropriate damage function depends on the user's input on the height and exposition of the building. An additional adjustment of the property-specific vulnerability is performed by respective factors considering e.g., roof type and construction year, which were provided by the *General Association of the German Insurance Industry* (*GDV*).

#### 3.4.1.2 Hail

Meteorological data on local hail hazard has been tracked systematically only for a relatively short-time period. Additionally, the correlation between hail and the expected damage to buildings is little studied, compared to storm or flooding. Determining the hail hazard of a location is challenging, since the number of local hail days is not sufficient for the desired purpose, as different degrees of hail intensity would be ignored. Empirical research identifies the kinetic energy of hailstones falling during a hail event as the key variable for modeling the damage to buildings (Schiesser et al., 1999; Hohl, 2001). This kinetic energy is depicted as hail intensity and depends mainly on the number and size of falling hailstones and the duration of the hailstorm. The strength of the hail event is captured by means of radar data, or precisely measurements of the so-called radar reflectivity, which is proportional to the size and number of hailstones. Radar reflectivity is measured as the logarithmic unit *dBZ*. For the *ImmoRisk* project, *KIT* provided reflectivity data for the years 2004-2011. The data are based on the so-called *RX product* of the 16 radar stations of the *DWD*. Data for future hazards

from climate models was not available within the timeframe of the project. For validation purposes, the data were filtered with lightning data to exclude erroneous radar echoes. Radar echoes were only considered, if lightning was also registered within a 5km radius. In order to apply the reflectivity data to the available damage functions, a functional relation between radar reflectivity and hail intensity is derived from data of Schiesser et al. (1999) for the *ImmoRisk* project. The applied damage functions are based on empirical work of Hohl (2001), who determined logistic damage functions for residential properties by combining meteorological data and observed damage cases. Further adjustments are conducted by applying building type specific factors provided by the *GDV*.

#### 3.4.1.3 Flood

The hazard from flood is specifically modeled by the *GFZ* for the locations covered by the *ImmoRisk* project. Modeling this hazard is more difficult than storm or heat, as the required variable water depth is not directly contained in climate models, as opposed to wind velocity or temperature. In contrast, the flood modeling requires a variety of complex models starting with rainfall from *dRCM COSMO-CLM*[13].

Rainfall serves as an input parameter for the hydrological model *SWIM*, which enable analyses of the statistical frequency of certain discharge levels ($m^3/s$) at individual locations in a modeled drainage network. Using the empirical relations between discharge and water level, as well as a three-dimensional, digital terrain model local water depths are finally determined. The used terrain model has a spatial resolution of 10m×10m and therefore allows a very precise estimation of flood hazards for a single property.

The vulnerability is modeled using the damage function *FLEMOps+* (flood loss estimation model for the private sector, see Buechele et al., 2006; Thieken et al., 2008). This is an empirical damage function derived and validated by representative interrogations and observations after flood events (see Figure 2). The used variables, that influence the resulting damage, are water depth outside the building (in five categories), the type of use, the standard of equipment and the existence of any private precautionary measures. Furthermore the existence of a cellar was considered according to empirical results of Kreibich and Mueller (2005).

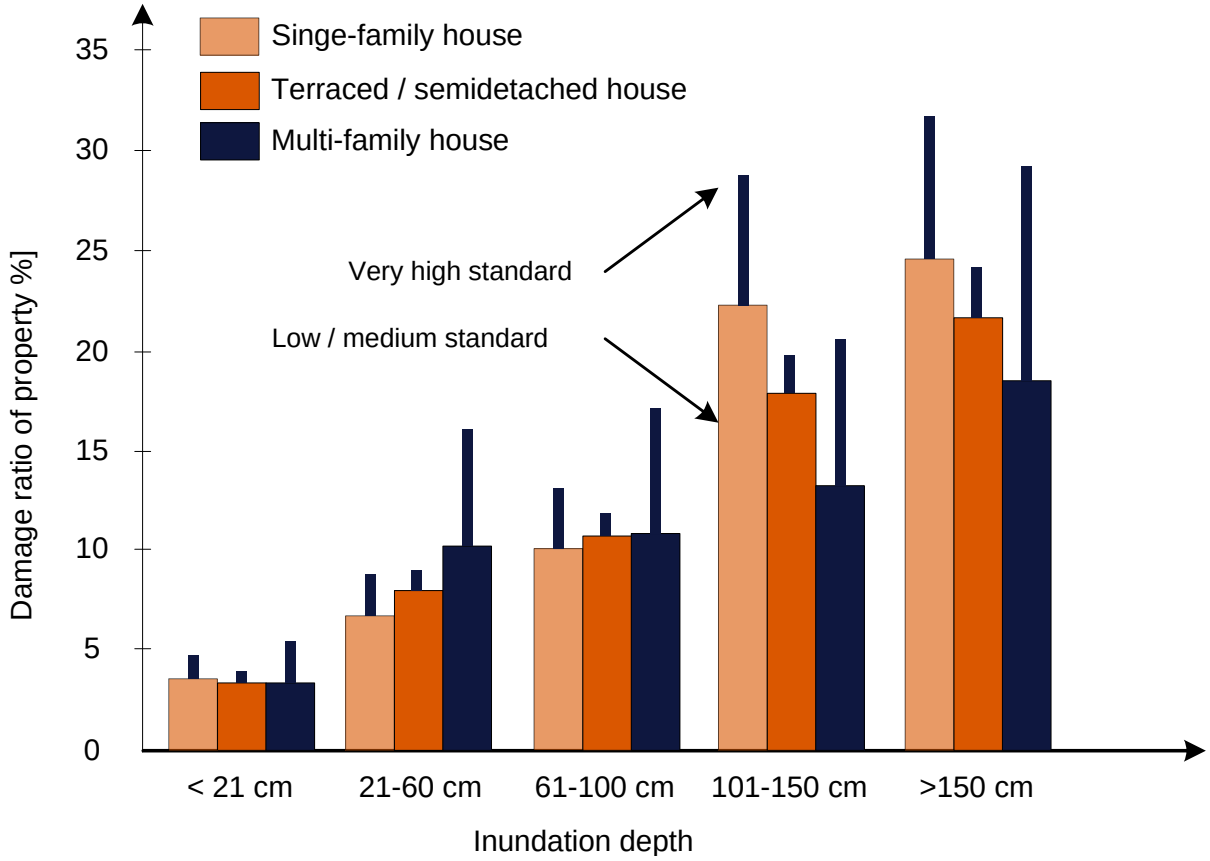
#### 3.4.1.4 Other Natural Hazards

For the other recorded natural hazards heat, heavy precipitation, forest fire and excess voltage, no damage functions were available during the project period. Therefore, information in the risk profile refers to the qualitative hazard of the location and does not consider property-specific vulnerability.

Information on the hazard of heat is based on data referring to the number of annual heat days at the respective location. Meteorologists define a heat day or also ‘hot day’ as a day with a maximum temperature of at least 30°C. As reference period for assessing the current hazard, the timeframe between 1970 and 2000 was selected. Trend information refers to the change of annual heat days between the reference period and the period 2071-2100. This future period was estimated using an ensemble of the three *RCM REMO, COSMO-CLM* and *WETTREG*. Data were provided by the *DWD*.

[13] Driven by *GCM ECHAM5* under the *A1B scenario* for the periods 1960-2000 and 2001-2100.

**Figure 2: FLEMOps damage functions**




Source: Based on Thieken et al. (2008)

The classification of hazard from heavy precipitation is based on data from *DWD*. The data were sourced from the system *KOSTRA-DWD-2000*[14] describing rainfall levels and duration of certain recurrence intervals. For example, in 100 years a location can experience 35mm rainfall within 10 minutes or 64mm within 60 minutes. On the one hand, heavy precipitation can directly damage properties, if, for example, gutters are not able to handle the prevailing forces. On the other hand, particularly urban areas with a high degree of sealed surface suffer from so-called flash floods, extensively draining off currents of water with significant depth. Hazard from these flash floods depends strongly on the local topography, but also on the dimensioning and reliability of drainage equipment. As these factors are not reflected in the *ImmoRisk* tool, the user should gather additional information on the location and property risk in case of increased meteorological hazard.

As most *ImmoRisk* sites are located in urban areas, the hazard from forest fire is estimated to be rather low. The so-called Forest Fire Index with hazard levels from 1 to 5 bases on values of precipitation, temperature and wind as the key parameters for the emergence and spreading of fires. The prediction of future hazard is based on the change of annual days featuring a hazard level of at least four and was derived from an ensemble of four regional climate models: *COSMO-CLM*, *REMO*, *WETTREG* and *STARS* (Wittich, 1998). Data were provided by the *Department for Agricultural Meteorology Research* of the *DWD*.

[14] Basis period 1951-2000

The hazard from thunderbolt bases on statistics of the regional frequency of cloud-ground lightnings. The hazard from excessive voltage does not result from the direct strike of lightning into a building, but from strikes in the neighborhood, which can damage electronic equipment through voltage surges. The hazard from excessive voltages does not necessarily correlate with the danger from lightning and is usually lower in urban areas than in rural areas. The data were provided by *GDV* (Table 1).

### 3.4.2 IT implementation

The *ImmoRisk* tool has been conceived as an online tool based on the client-server model. All user requests are registered and processed on the server and the results of the risk calculations are sent to the user. The advantage for the user is that downloading the software is unnecessary, as well as the download of updates. Additionally, sophisticated calculations can be performed by the server, so that the requirements of the user's hardware are comparatively low. From an *IT* perspective, an online tool is advantageous, as there is no need to program several different versions for all established operating systems. Only small adjustments to partially divergent ways of presentation between the browsers had to be considered.

**Table 1: Summary of data used in the ImmoRisk tool**

| Climatic Aspect | Hazard | Vulnerability |
|---|---|---|
| Storm | Extreme value statistics from *CCLM* (4 runs) driven by *ECHAM5-OM* & *CCCma3*, two periods (*KIT*) | Property-specific damage functions and adjustment factors (insurance industry) |
| Hail | Thunderbolt-filtered radar reflectivity, *RX product* (*DWD* and *KIT*); application of extreme value statistics by the authors | Empirical damage functions and property-specific adjustment factors (Hohl, 2001 and insurance industry) |
| Flood | Location-specific modeling on flood hazards, rainfalls from *ECHAM5-CLM*, two periods (*GFZ*) | Empirical damage functions *FLEMOps+* from current research (Thieken et al., 2008, Mueller and Kreibich, 2005) |
| Heat | Number of heat days from Regional Climate Model ensemble, *REMO*, *CCLM*, *WETTREG*, two periods (*DWD*) | --- |
| Forest fire | Current hazard and trend from Climate Model ensemble, *RCM REMO*, *CCLM*, *WETTREG*, *STARS* (*DWD*) | --- |
| Heavy precipitation | Recurrence intervals of heavy rainfall events from *KOSTRA-DWD-2000* (*DWD*) | --- |
| Thunderbolt, excessive voltage | Hazard data (insurance industry) | --- |

The basic framework of the webpage is written in *HTML* using *Cascading style sheets* (*CSS*) in order to determine general layout of the pages. *Hypertext Preprocessor* (*PHP*) is used to handle user input, conduct calculations and create the *PDF*-version of risk profiles. *JavaScript* facilitates dynamic elements like filters and interactive help inserts. The interactive drag-and-drop selection of a building's micro-location is implemented with *Asynchronous JavaScript and XML* (Ajax) and some *JavaScript* template from *script.aculo.us* (http://script.aculo.us/). The calculation of *AEL* data are conducted with the proprietary software package *GAUSS* (*Aptech*), which is controlled via *PHP*.

## 3.5 Functional implementation of the ImmoRisk tool

The implementation of a tool which meets the abovementioned requirements of the real estate industry is realized as an online tool, which is accessible via web browser.

A user account system was developed in order to enable users to record their own portfolios and to adjust them at any time. After signing up for the tool, the user can access a sample portfolio, partially based on actual buildings at the selected pilot locations. In this way, the user can get used to the results of a risk assessment before entering and analyzing the own properties. The *ImmoRisk* tool allows a multitude of residential properties at the available locations in order to compare different construction features of the buildings. In addition, the web tool offers a contact form for sending questions to the developers and offers a user manual and a document as downloads, which provides further information about content basics and the data.

In a first step, the user defines the macro and micro-location of the property (see Figure 3). After selecting one of the 15 pilot locations (see Figure 4), the user can readjust the exact position by drag-and-drop onto an aerial image. In particular, the selected location is crucial for the hazard, and influences the regional factor for valuation purposes. Defining the micro-location plays an important role in determining the hazard caused by flood, which varies considerably on a small scale. In contrast, the hazard for all other natural disasters is defined by the macro-location.

In the second step (see Figure 5), the user defines a set of constructional features that are applied for calculating the vulnerability and the property value. The recorded features comprise, among others, the type of use (e.g. detached house, terraced house, multi-family house), the construction type, the roof shape, building age, height, gross floor area, basement area, renovations and special exposition (e.g. resulting from a location in an open area).

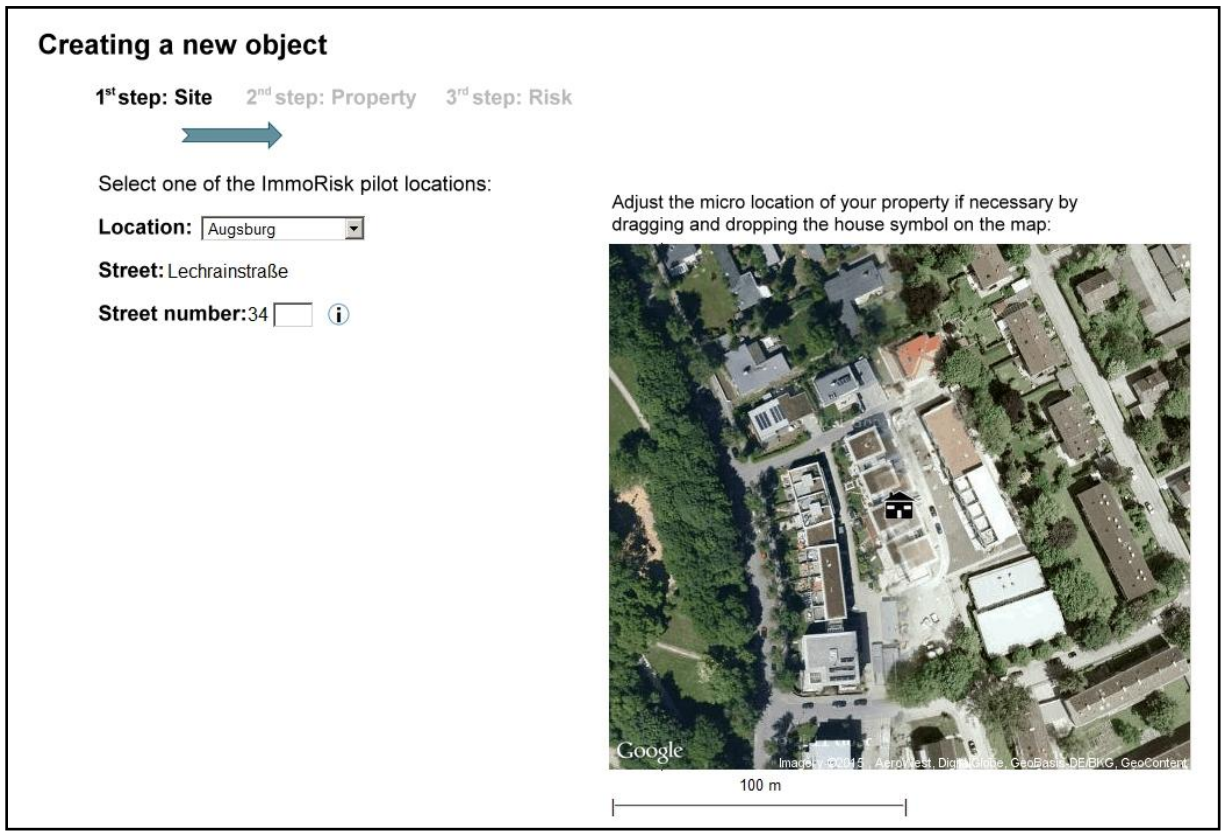
When all input data are entered, the property is saved in the user's portfolio and a risk profile is displayed, which can automatically be converted into a downloadable *PDF* file. The risk profile contains a summary of all essential data on the property, as provided by the user, as well as all results of the hazard and risk analysis (see Figure 6). As valid damage functions were available only for storm, hail and flood, a quantitative risk analysis resulting in an *AEL* was only convertible for these natural risks. In addition, no data on future hazards were available for hail at the time of implementation. *AEL* is complemented by the so-called damage rate, which is customary in the insurance industry and equals the *AEL* in relation to the corresponding property value. The information on the average AEL is accompanied by respective margins of uncertainty, which are described by the upward and downward 95 percent quantile.

Furthermore, the tool provides qualitative information on a five-level scale with respect to the location-specific hazard and its expected shift as a result of climate change for the following natural hazards: storm, flooding, hail, heat, heavy precipitation, forest fire, lightning and excess voltage.

All properties entered by the user are saved in his portfolio (see Figure 7). In this portfolio view, the essential data with regard to the hazards of individual properties are visible. Additionally, the user can retrieve risk profiles, delete or copy properties, change information about location or constructional features, filter or sort their properties by location and view the buildings on a map.

By varying the macro or micro-location or constructional features, the user can validate the influence of his potential investment decisions on the quantitative risk caused by natural hazards and generate the respective action-relevant information.

**Figure 3: Menu for the site selection in the ImmoRisk tool**




*Source: ImmoRisk tool*

**Figure 4: Locations covered by the ImmoRisk tool**

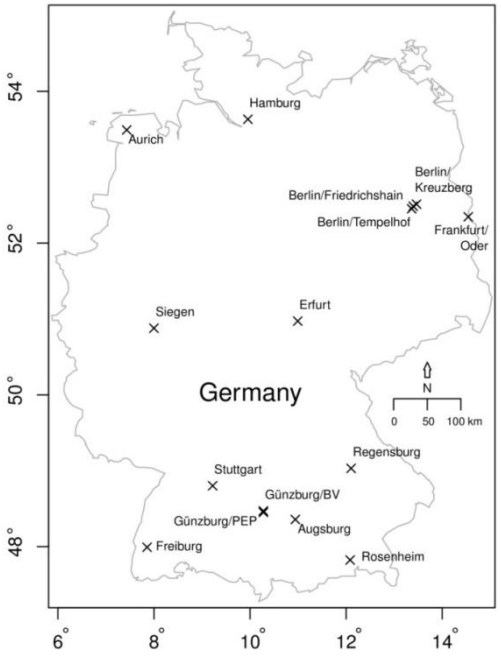
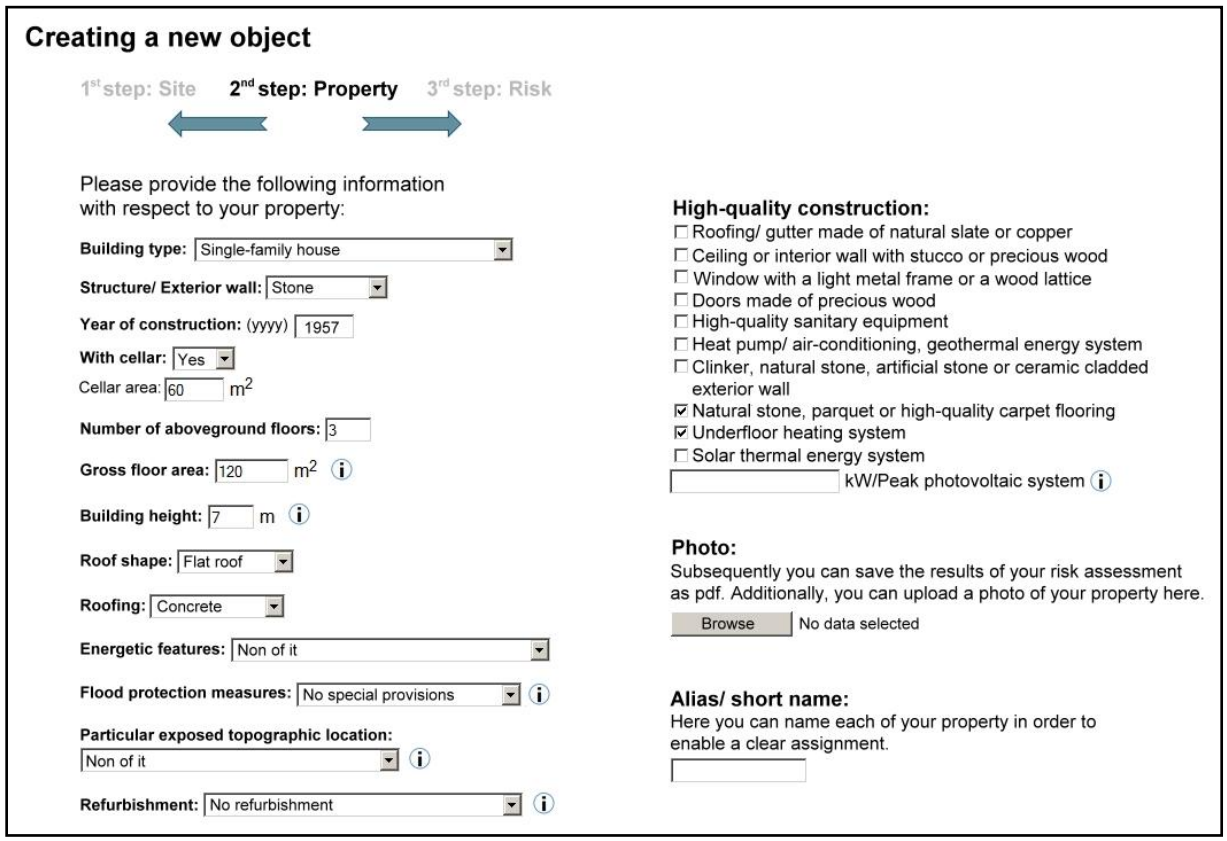
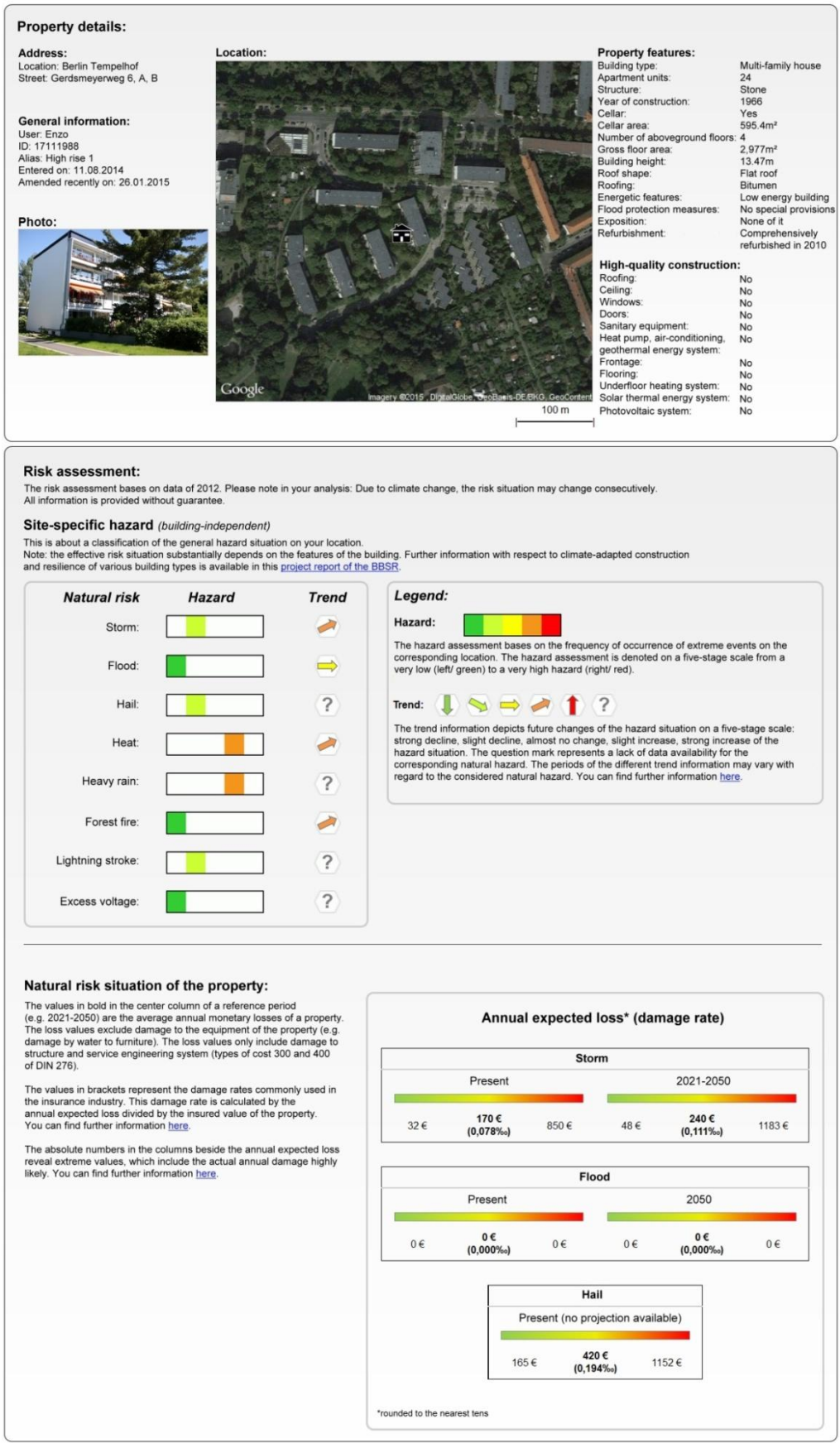


**Figure 5: Menu for the selection of building specific feature in the ImmoRisk tool**




*Source: ImmoRisk tool*

**Figure 6: Risk profile generated by the ImmoRisk tool**




*Source: ImmoRisk tool*

**Figure 7: Portfolio menu of the ImmoRisk tool**

| Property-ID (assorted) | Location (assort) | Street | Alias | Photo | Hazard: Storm | Flood | Hail | Heat | Forest fire | Heavy rain | Lightning stroke | Excess voltage | Entered (amended recently) | Delete | Copy | Open profile |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Property#4: | Stuttgart | Dürrheimer Straße | Pavillon |  | | | | | | | | | 14.07.2012 (22.04.2014) | | | |
| Property#7: | Berlin Tempelhof | Gerdsmeyerweg 6,A,B | |  | | | | | | | | | 14.07.2012 (20.04.2015) | | | |
| Property#8: | Stuttgart | Wildunger Straße 14 | |  | | | | | | | | | 14.07.2012 (21.09.2012) | | | |
| Property#11: | Berlin Tempelhof | Gerdsmeyerweg 8 | | | | | | | | | | | 09.08.2012 (20.09.2012) | | | |

*Source: ImmoRisk tool*

## 3.6 Risk assessment results of the ImmoRisk pilot study

The subsequent section deals with the assessed monetary risks (available for storm, hail and flood) and does not take account of those natural disasters that are only analyzed qualitatively without considering property-specific vulnerability (heat, forest fire, heavy precipitation, thunderbolt and excessive voltage). Due the experimental nature of the project with only 15 pilot locations the results have to be regarded as preliminary.

### 3.6.1 Storm

Risk calculations by the *ImmoRisk* tool demonstrate that there is large cross-sectional variation between the 15 locations (see Figure 8). For reasons of comparability, a standard building (two storey single-family house, construction year 2008, flat roof, 157sqm gross floor area) is defined in order to conduct risk calculations for each location. The consideration of damage rates, instead of *AEL*, prevents the bias of regional diverging construction costs that affect the *AEL*. Storm damage rates range from 0.074‰ for Regensburg in southern Germany to 0.234‰ for Erfurt in central Germany, under present climate conditions. Assuming approximate replacement costs of € 1,200 per square meter, these damage rates correspond to an *AEL* from storm damage of €14 and €44, respectively. These data were cross-checked with corresponding insurance premiums of a comparable building and location. It has been shown that the calculated *AEL* of the *ImmoRisk* tool are within a realistic range and describe local hazard variations in a more detailed manner than insurance premiums. The mean damage rate for storm amounts to 0.117%, corresponding to an *AEL* of €22. Considering the period 2021-2050, there is evidence of a rising risk caused by storm events at all locations. The mean damage rate rises by approximately 40% to 0.164‰ (€31). The highest increase in damage rates is identified in the coastal areas of northern Germany, followed by eastern and central Germany. The lowest increase is observed in Freiburg, southwest Germany, with 3.8%.

### 3.6.2 Hail

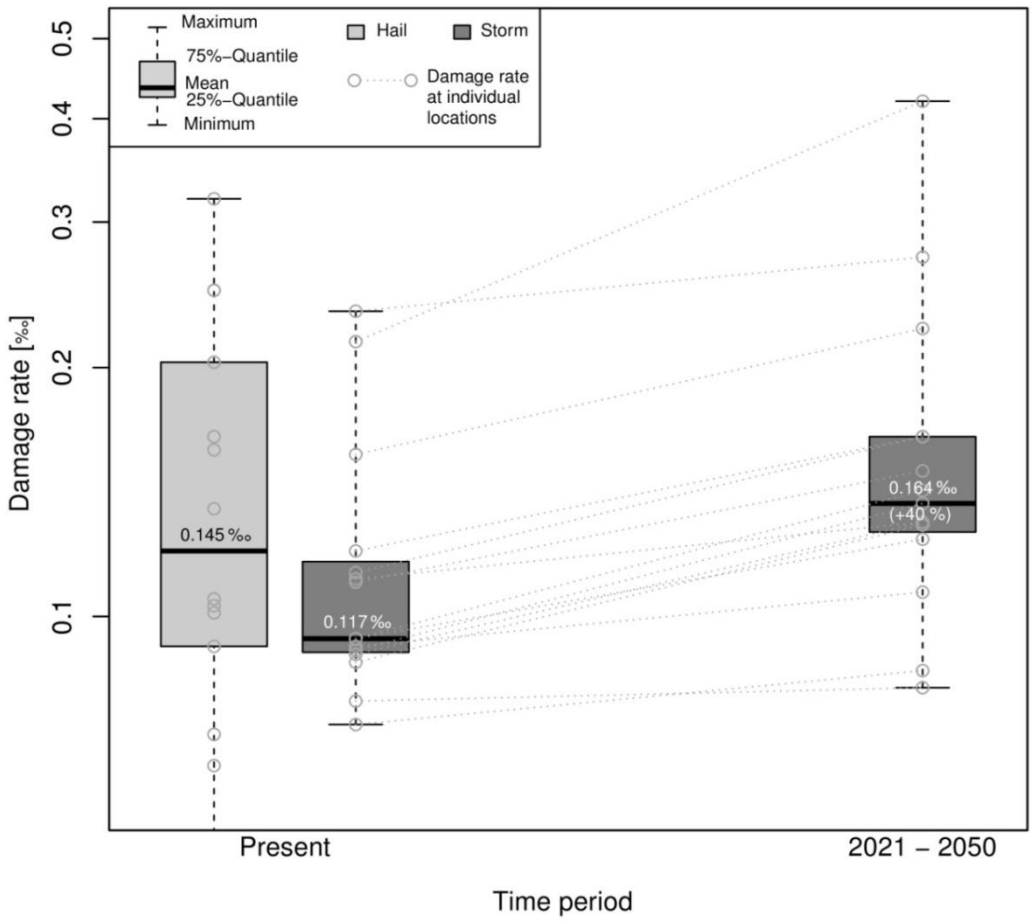
The reverse picture emerges when analyzing the local hail risk by the *ImmoRisk* tool. The highest values of damage rate are revealed in southwest Germany (Stuttgart 0.320‰, Freiburg 0.203‰), whereas the risk in the coastal areas of northern and eastern Germany is quite low (Aurich 0.013‰, Berlin/Kreuzberg 0.066‰). The mean damage rate for hail amounts to 0.145‰ and is therefore

higher than for storm risk. As in the case of storm, these calculated values are in the range of real life insurance premiums. Anyway, our results feature a distinctly higher spatial variation in comparison to the evaluated insurance premiums and facilitate a more realistic view of the actual risk situation. The combined risk of storm and hail amounts to a mean of approximately 0.254‰ (in deviation from the sum of mean hail and storm damage, because no hail hazard data were available at one location, which was excluded therefore), corresponding to an *AEL* of almost €48 per year for the investigated standard building. As no data on future hail hazard were available, we could not identify the effects of climate change on hail risk.

### 3.6.3 Flood

Flood risk is only relevant for three of the investigated pilot locations. The other locations are not exposed to flood risk at all. Within the locations exposed to flood risk, damage rates vary considerably depending on the micro-location of the building. Thus, it is not possible to make general statements about the flood risk of the 15 pilot locations. The highest values are found at Guenzburg/*PEP* with damage rates above 3‰ for present and future climate conditions. The damage rates in Stuttgart are significantly lower but still considerable, at least on parts of the investigated site. The pilot location Augsburg presents no flooding in the median simulation data. Only considering the 95% confidence quantile, the data exhibit any flooding at all. Even considering this upper extreme, the damages are quite low according to minor inundation depths.

**Figure 8: Distribution of damage rates for storm and hail**




*Source: Calculated with the ImmoRisk tool by the authors*

## 3.7 Conclusion

Damage caused by natural hazards is globally increasing. In Germany, the majority of damage results from extreme weather events, in particular storms, flood and hailstorms. A causal correlation with the emerging climate change is assumed by climatologists. Therefore, climate change mitigation strategies are accompanied by adaptation strategies. Nevertheless, a sound information basis is needed, regardless of which stakeholder is considered, like policy-makers, private households, large scale or SME. In summary, this paper and its underlying tool *ImmoRisk* have been developed to provide a useful first step toward a platform assessing climate risks and contribute to the literature in four respects.[15]

First, whereas a variety of information and models is available providing data on the present and future climate, these platforms are insufficiently suitable to the requirements of the housing and real estate industry. Action-relevant data with respect to the specific risk of damage to properties are lacking. Most existing approaches tend rather to consider the point of view of insurances, instead of the real estate sector. In this case, the latter is aspired. Hence, a trinomial approach is used for the *ImmoRisk* tool comprising the natural hazard, the vulnerability and the value of the property.

Second, most of the existing tools do not consider possible monetary losses adequately, as the damages are calculated as an average value of an average building in an aggregated region like the zip code level. However, in the case of investing in real estate or the installation of adaptation measures, investors require information on the risk associated with the specific property, ideally considering risk-relevant constructional features. New regional climate models improve the hazard data situation, which calls for model specification with regard to the vulnerability and the value component beyond top-down approaches. To the best of our knowledge, *ImmoRisk* is the first publicly available tool with respect to climate risks which uses a property-specific bottom-up approach. This approach could be used in future research, to gain a better understanding of extreme events with high impact and low probability. The so-called black swan events constitute a considerable challenge for property managers (see Higgins, 2014) and can be analyzed with a framework that is analogue to the one applied in the *ImmoRisk* tool.

Third, most existing platforms address the current climate only and do not provide information on the adjusted risk situation caused by climate change. Their output does not conform to the long-term requirements of real estate risk management. Long-term investment decisions require a comprehensive data basis with respect to climate forecasts of the most sophisticated climate models, as locations are fixed for a long period. *ImmoRisk* uses state of the art hazard data including forecasts of regional climate models. Even if data on future hail hazard was not available during the project period, the *ImmoRisk* project demonstrated how such data can be processed in the future.

Fourth, for real estate risk management purposes, it is essential to cover all relevant natural hazards which exert a significant influence on the risk situation of a property in the study area. Accordingly, an appropriate tool for identifying natural risks in the German-speaking area must consider storms,

[15] The information required for this kind of risk assessment already exists in part, is not concentrated at one institution and thus cannot be used without sophisticated linkage.

flooding, as well as hailstorms. In contrast to the *ImmoRisk* tool, existing approaches lack a simultaneous consideration of the natural risks or do not consider hailstorms at all.[16]

The research agenda within real estate risk management with regard to climate risks is still long. In general, a more efficient coordination between climate research institutions and the requirements of risk research is highly recommended. In particular, research should focus on an expansion of the tool from 15 pilot locations to a nationwide geographic information system, which could be developed based on the trinomial *ImmoRisk* approach comprising hazard, vulnerability and value components. Originally launched as an experimental project, *ImmoRisk* demonstrates the feasibility of a complex natural hazard information system and offers distinct links for further advances in the transparency of future climate risks in the real estate sector. Further research could focus on additional building components or could address interactions between simultaneous extreme weather events like storm and heavy precipitation which boost damages.

Beside these recommendations for researchers, the lessons learned for different stakeholders are as follows. Both for private households and for companies, the *ImmoRisk* tool provides a suitable basis for investment decision-making regarding the choice of location, individual adaptation measures or weighing up insurance costs with accepted future damage. Whereas large-scale companies can eliminate parts of their location risk according to their portfolio diversification strategy, the site selection of private households or *SME* experiences difficulties, as their properties and particularly their sites are fixed in the medium run. But even institutional investors need reliable information about increasing natural hazards and their impact on assets. The contribution for planning and construction practice is manifold. The *ImmoRisk* tool provides a valuable basis for assessing the economic efficiency of extensive adaptation activities like flood protection measures, refining existing or redeveloping new instruments of construction and planning practice, providing appropriate guidance for future construction and material standardization or designating building land.

## 3.8 References

André, C., Monfort, D., Bouzit, M., Vinchon, C. (2013). Contribution of insurance data to cost assessment of coastal flood damage to residential buildings: insights gained from Johanna (2008) and Xynthia (2010) storm events, *Natural Hazards and Earth System Sciences*, 13(8), 2003-2012.

Bienert, S., Hirsch, J., Braun, T. (2013). *ImmoRisk – Risikoabschaetzung der zukuenftigen Klimafolgen in der Immobilien- und Wohnungswirtschaft*, in: BMVBS (Ed.), No. 159, BMVBS, Forschungen, Bonn, 1-101.

Buechele, B., Kreibich, H., Kron, A., Thieken, A., Ihringer, J., Oberle, P., Merz, B., Nestmann, F. (2006). Flood-risk mapping: contributions towards an enhanced assessment of extreme events and associated risks, *Natural Hazards and Earth System Sciences*, 6(4), 485-503.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*, Springer series in statistics, London, New York.

[16] Hailstorms in 2013 caused damage of approximately € 3.1 billion to the property insurers, which considerably exceeded the damage of storms and flooding. As just parts of the damage were covered, the actual economic costs are assumed to be distinctly higher.

Friedland, C. J. (2009). *Residential building damage from hurricane storm surge: proposed methodologies to describe, assess and model building damage*, Dissertation, Louisiana State University, Baton Rouge.

GDV (Ed., 2014). Naturgewalten 2013 – Sieben Milliarden Euro zahlten Versicherer fuer Hochwasser, Stuerme und Hagel, online: http://www.gdv.de/2014/01/sieben-milliarden-euro-zahlten-versicherer-fuer-hochwasser-stuerme-und-hagel/ (accessed: 23.04.2014).

German Federal Government (Ed., 2008). German Strategy for Adaptation to Climate Change, online: http://www.bmub.bund.de/fileadmin/bmu-import/files/english/pdf/application/pdf/das_gesamt_en_bf.pdf (accessed: 23.04.2014).

Heneka, P., Hofherr, T., Ruck, B., Kottmeier, C. (2006). Winter storm risk of residential structures – a model development and application to the German state of Baden-Wuerttemberg, *Natural Hazards and Earth System Sciences*, 6(6), 721-733.

Higgins, D. M. (2014). Fires, floods and financial meltdowns: black swan events and property asset management, *Property Management*, 32(3), 241-255.

Hohl, R. (2001). *Relationship between Hailfall Intensity and Hail Damage on Ground, determined by Radar and Lightning Observations*, Dissertation, University of Fribourg.

IPCC (Ed., 2013). *Climate Change 2013 – The Physical Science Basis*, Working Group I Contribution to the 5th Assessment Report of the IPCC, IPCC, Geneva.

ImmoRisk tool. Federal Ministry for Transport, Building and Urban Development (Ed., 2013), online: http://xrl.us/immorisk (accessed: 23.04.2014).

Jongman, B., Kreibich, H., Apel, H., Barredo, J. I., Bates, P. D., Feyen, L., Gericke, A., Neal, J., Aerts, J. C. J. H., Ward, P. J. (2012). Comparative flood damage model assessment: towards a European approach, *Natural Hazards and Earth System Sciences*, 12(12), 3733-3752.

Kleiber, W., Fischer, R., Simon, J. (2010). *Verkehrswertermittlung von Grundstuecken – Kommentar und Handbuch zur Ermittlung von Marktwerten (Verkehrswerten), Versicherungs- und Beleihungswerten unter Berueecksichtigung der ImmoWertV*, 6th edition, Bundesanzeiger, Cologne.

Kleist, L., Thieken, A. H., Koehler, P., Mueller, M., Seifert, I., Borst, D., Werner, U. (2006). Estimation of the regional stock of residential buildings as a basis for a comparative risk assessment in Germany, *Natural Hazards and Earth System Sciences*, 6(4), 541–552, online: http://www.nat-hazards-earth-syst-sci.net/6/541/2006/ (accessed: 23.04.2014).

Kunz, M., Puskeiler, M. (2010). High-resolution assessment of the hail hazard over complex terrain from radar and insurance data, *Meteorologische Zeitschrift*, 19(5), 427-439.

Markose, S., Alentorn, A. (2011). The Generalized Extreme Value (GEV) Distribution, Implied Tail Index and Option Pricing, *The Journal of Derivates*, 18(3), 35-60.

Merz, B., Kreibich, H., Thieken, A., Schmidtke, R. (2004). Estimation uncertainty of direct monetary flood damage to buildings, *Natural Hazards and Earth System Sciences*, 4(1), 153-163.

Mueller, M., Kreibich, H. (2005). Private Vorsorgemassnahmen koennen Hochwasserschaeden reduzieren – Nutzung der Kellerraeume beeinflusst die Schadenhoehe, *Schadenprisma: Zeitschrift fuer Schadenverhuetung und Schadenforschung der oeffentlichen Versicherer*, 2005(1), 4-11.

MunichRe, Munich Reinsurance Company, Geo Risks Research, NatCatSERVICE, Munich (Ed., 2014). Raw data received directly from Munich Re.

Naumann, T., Nikolowski, J., Golz, S. (2009). *Synthetic depth-damage functions – a detailed tool for analysing flood resilience of building types*, in: Pasche, E., Evelpidou, N., Zevenbergen, C., Ashley, R., Garvin, S. (Eds., 2008). Road map towards a flood resilient urban environment, Proceedings of the

international Conference of the COST action C22 urban Flood Management in Cooperation with UNESCO-IHP, Hamburger Wasserbauschriften, Vol. 6, 16-23.

RICS (Ed., 2009). RICS Valuation Standards (The Red Book), 6th edition, London.

Sauer, M. (2010). *Vergleich der Klimasimulationsergebnisse der Regionalmodelle CLM, REMO, STAR II und WETTREG fuer Hannover – Nr. 1: Modelleigenschaften und –unterschiede*, Hannover.

Schiesser, H.-H., Hohl, R., Schmid, W. (1999). *Ueber die Beziehung Hagelfall – Gebaeudeschaeden: Fallstudie 'Luzern-Hagelsturm' vom 21. Juli 1998*, Zurich.

Thieken, A. H., Mueller, M., Kleist, L., Seifert, I., Borst, D., Werner, U. (2006). Regionalisation of asset values for risk analyses, *Natural Hazards and Earth System Sciences*, 6(2), 167-178, online: http://www.nat-hazards-earth-syst-sci.net/6/167/2006/nhess-6-167-2006.pdf (accessed: 23.04. 2014).

Thieken, A., Olschewski, A., Kreibich, H., Kobsch, S., Merz, B. (2008). *Development and evaluation of FLEMOps - a new Flood Loss Estimation MOdel for the private sector*, in: Proverbs, D., Brebbia, C. A., Penning-Rowsell, E. C. (Eds., 2008). Flood recovery, innovation and response I, WIT transactions on ecology and the environment, Vol. 118, 315-324, WIT, Southampton.

Unanwa, C., McDonald, J., Mehta, K., Smith, D. (2000). The development of wind damage bands for buildings, *Journal of Wind Engineering & Industrial Aerodynamics*, 84(1), 119-149.

Wittich, K.-P. (1998). Waldbrandgefahren – Vorhersage des Deutschen Wetterdienstes, *AFZ/Der Wald*, 53(6), 321-324.

## 3.9 Appendix

**Climate projection:** A climate projection is referred to as potential future development of individual or several climate parameters, calculated based on scenarios by a climate model.

**Damage rate:** The damage rate is an actuarial value and equals the *AEL* in relation to the corresponding property value.

**Digital terrain model (DTM):** A *DGM* is a database for describing the relief of an area in order to derive contour maps, volumes and slopes.

**Exceedance Probability (PE):** Probability for exceeding a certain value $h_{krit}$ (e.g. runoff, amount of precipitation, wind speed) in a given period: $PE = P\ (h > h_{krit})$. Example: A centenary event corresponds with an annual probability of exceedance of 1%. Incidentally, the probability of the actual occurrence of a centenary flood in a period of 100 years amounts 63%.

**Exposition:** The exposition as part of the vulnerability depicts in the context of climate impact research, to what extent a region or system is faced with considered changes of the climate parameters (e.g. temperature, precipitation).

**Extreme value statistics – Frequency Analysis:** Statistical analysis of extreme values in order to determine the exceeding probability of certain values. In the fields of hydrology and meteorology the annual maximal values commonly serve as a basis for extreme value statistics.

**Extreme weather events:** An extreme weather event is an event associated with extreme weather conditions such as heat, storm or heavy precipitation, which is seldom for the given location and season.

**Gross floor area (sqm):** The gross floor area is defined by the aggregated floor area of all floor levels of a property.

**Hazard:** A potentially harmful physical event, phenomenon and/ or human activity, which can lead to a loss of human life or injuries, property damages, social and economic disruptions or environmental destructions. It is independent of the type of potentially affected human beings or objects. Therefore, it only describes the event-specific hazard. Hazards may include latent conditions which may represent future threats and may have different origins: natural (geologic, hydrometeorologic and biologic) and/or caused by human activities (environmental destruction and technologic hazards). They may occur individually, subsequently or simultaneously depending on their origins and impacts. Each hazard is characterized by its location, intensity, frequency and probability. In the context of hazard- and/or risk assessments, a natural hazard is defined by the probability of occurrence of a natural phenomenon associated with catastrophe potential in a given area and period.

**Risk:** This term has many meanings:

(1) Generally, risk is the possibility to suffer damage caused by a hazard.
(2) In the context of the insurance industry, an object to be insured is referred to as risk.
(3) In risk sciences, the term risk encompasses the probability and the amount of harmful consequences or expected losses resulting from interactions between natural or human induced hazards and vulnerable conditions. The quantification of risk can focus on selected risk elements and/or on one or several damage types. For example, the following damage indicators may be assessed: number of dead persons, injured persons and/or evacuees, economic reconstruction or time value of the destroyed assets (infrastructure, buildings, contents, machines, inventories, vehicles, agricultural and silvicultural products etc.), value of indirect economic damage (caused by interruptions of operations and services, loss of financial assets etc.) and sociocultural damage as a result of damage of cultural artifacts, livelihoods (e.g. contaminated water or soil), environmental and scenically valuable regions.

**Runoff, discharge:** The part of the fallen precipitation, which runs off in creeks and rivers (underground included) measured in $m^3/s$: The direct runoff includes surface runoff and underground interflow, whereas total runoff additionally includes ground water base flow.

**Scenarios:** Scenarios are coherent, consistent and plausible descriptions of potential future conditions including the development of their emergence. They base on assumptions. The descriptions may proceed qualitative or quantitative.

**Spatial resolution:** A measure of the smallest identifiable area on a picture as a discrete independent unit. The spatial resolution describes the size of the smallest describable object of a digital data set. The term is generally used in terms of grid data models. The resolution of a raster corresponds with the size of the cell of the real world. Accordingly, the resolution of a satellite image may be 10m (i.e. each pixel represents 10mx10m of the ground).

# 4 The analysis of customer density, tenant placement and coupling inside a shopping centre

**Jens Hirsch** **Matthias Segerer** **Kurt Klein** **Thomas Wiegelmann**

**Abstract**

The spatial arrangement of tenants is currently one of the main topics in shopping centre research. This paper shows how a *Geographic Information System* (*GIS*) can be used to analyse the tenant structure. Given the recommendations in the literature, the analysis may help to improve the situation within a certain shopping centre. Therefore, we introduce the variable clumping method and kernel density estimation into shopping centre research in order to analyse retail category concentrations, customer flows and coupling in a shopping centre. Applying these techniques to a German shopping centre showed that spatial concentration can be observed within the retail categories of 'food, health & body and fashion' and that the pass ratio declines according to the distance from the central point of the shopping centre. Also, shops in the same retail category have higher coupling than those of different categories, and unexpectedly spatially separated shops have a slightly higher coupling than non-spatially separated ones. Overall, the use of *GIS* improves the quality and the speed of spatially based analysis, and thus should be used more frequently in scientific shopping centre research and shopping centre management.

## 4.1 Introduction

The success of real estate investment is largely based on location and this applies specifically to retail assets (Levy et al., 2013). Surprisingly, the use of techniques which can address spatial issues is still not very popular within real estate research and management. Dubin et al. (1999) comment on this situation as follows:

> *"Ironically, real estate as a discipline espouses the supremacy of location while employing economic tools designed for a spaceless world. Adoption of spatial statistical techniques offers the opportunity to align theoretical considerations with empirical practice."*

Against a background of increasing competition with providers of online shopping and customers who are better informed, more sovereign and less loyal (Stepper, 2014), the operators of shopping centres (*SC*) should make use of innovative technical solutions in order to improve their business. Nowadays, Geographic Information Systems (*GIS*) have become accepted as a data handling and analytical tool in the real estate industry and in real estate research (e.g. Castle, 1998; Culley, 2010; Klein, 2007; Segerer, 2011; Thrall, 1998). In this context, these papers focus on the analysis of real estate locations from an external perspective, but a change in perspective is now occurring regarding the use of *GIS*-applications. *GIS* is not only used 'around real estate', but within it. In this respect, *SC* are one of the types of real estate which - due to the optimisation of passers-by-frequency and tenant selection/ arrangement - lend themselves to more detailed examination. Yuo (2010) was one of the first to use *GIS* as an analytical and optimisation tool within a shopping centre by focusing on tenant distribution and visitor control. These topics involve two of the main problems of shopping centre management, namely the optimisation of tenant selection and the distribution of shops. Based on this spatial problem, one aim of this study is to demonstrate how a broad range of objective information about the structures within a shopping centre can be obtained from *GIS*-analysis.

Given this focus on shopping centre research, the paper is structured as follows. Section 4.2 provides an overview of literature on tenant mix that is relevant to this research. Section 4.3 outlines possible research fields in terms of *GIS*-use and tenant arrangements, presenting research data and conducting *GIS*-analysis of the *SC*. Based on this analysis, the ideal tenant arrangement is discussed in Section 4.4. The paper concludes with the main results and questions for future research.

## 4.2 Literature review: shopping centre tenant mix

Tenant selection and shop arrangements have been viewed from various theoretical perspectives. Potential tenants expect a bid mix, which enables them to take advantage of synergetic effects (including shared and suspicious businesses) (Nelson, 1958) and which can minimise internal competition - in order to make the competition tolerable. Similarly, the consumer needs to minimise his or her transaction costs in the *SC* - as long as it is a matter of a rationally organised visit. Under these circumstances, customers should prefer clustered arrangements, which minimise the effort of visiting all shops in a specific retail category. If consumers prefer the adventure, the amenity value will have a greater impact on the arrangement of the shops. Yuo and Lizieri (2013) demonstrated that the decision between clustering and departmentalisation depends on physical features of the centre, namely the number the storeys and the spatial complexity of the floor plan. More storeys and

greater complexity lead to higher transaction costs for customers and therefore require an increased clustering of shops from the same retail category.

In the end, it is the *SC*-operator who has to reconcile the demands, and his optimisation criterion is the rent per sqm. According to Dawson (1983), shopping centre tenant mix considerations deal mainly with two key issues:

- the number, nature and size of tenants' outlets within the centre
- the placement of these outlets relative to each other and the points of entry into the complex

### 4.2.1 Tenant selection

Yuo et al. (2004) identify six factors which can be used for classifying shopping centre tenants by means of factor analysis. Based on this classification, they explain that the concentration of 'core' retail categories - such as Fashion, Comparison Variety, Selective Information and Health - raises shopping centre rents. Beside this positive correlation, the size of the shopping centre, number of shops, average size of shops and the number of categories and brands are positively related to shopping centre rent. Des Rosiers et al. (2009) have also dealt with the topic of retail concentration within a shopping centre. Like Yuo et al. (2004), they use the *Herfindahl index* to measure the (non-spatial) concentration of retail categories and demonstrate a rather negative correlation between intra-category retail concentration and rent. At the same time, the importance of 'higher retail categories' for rent optimisation was confirmed.

### 4.2.2 Tenant placement

Brown (1992) attempts to integrate spatial aspects into his research by assessing the 'passbuy ratio' and the coupling of shopping centre consumers in Belfast. Coupling denotes a visit to two specific shops during one single shopping trip. His findings are that - from a methodological perspective - this approach makes it possible to identify 'killing grounds' and - from a content perspective - this approach […] is a striking confirmation of the 'match' as opposed to 'mix' school of tenant placement. Shopping centre research that focuses on behavioural aspects, like *foraging theories* (see Wells, 2012), basically considers spatial aspects of customer behaviour, but does not explicitly take into account the role of retail categories. The general placement of stores from the same or different retail categories has not yet been incorporated into these models. This also applies to the promising efforts at modelling customer behaviour with multi-agent simulation (Horni et al., 2012; Siebers and Aickelin, 2007; Walid and Moulin, 2006).

Carter and Haloupek (2000) focus on improvements in data processing and modelling. They explain the nature of shopping centre rents by implementing spatial variables in their model and determining for every shop in the centre, its distance from the nearest exit, its distance from the nearest similar store, its distance from a central point, and its distance from the nearest vacant store. While the first two distance measures have no significant impact on rents, the distance from the nearest vacancy is significant at the 10% level and the distance from the centre at a 5% level. In addition, they focus on spatial autocorrelation by studying the spatial distribution of the residuals and state that the use of a spatial estimator improves the overall model fit by about 5%. Also, Carter and Haloupek (2002) explain that non-anchor stores of the same retail category should be dispersed within the shopping

centre by using a *p-median model*. Carter and Vandell (2005) investigate – using the distance measures of Carter and Haloupek (2000) – the arrangement of different store types based on the *bid rent theory*. Thus they confirm the general idea of the bid rent theory namely that different branches have a different 'ability to pay' (depending on turnover and profit margin ratio) and that the rent level declines from the centre to the edge of a shopping centre.

Analogously, the average unit size of shops decreases with their proximity to the mall centre. According to Carter and Allen (2012) the location of store categories obeys the laws of bid rent theory and revised central place theory. Therefore, stores with the steepest bid rent curve (in their case jewellery stores) will locate nearest the centre of the mall.

## 4.3 GIS-techniques

Achieving the ideal tenant placement system is the main task of shopping centre management. The literature specifies a number of optimisation criteria and this present study demonstrates how these criteria can be analysed with *GIS*, revealing potential for optimisation. The paper considers the substantial capabilities of *GIS* with respect to customer flow and tenant placement, especially regarding retail category concentration. By combining a customer survey and *GIS*-techniques, we further analyse some hypothesises in the recent literature. By applying a large customer survey, a database of spatial and non-spatial data was created, facilitating in-depth analysis. The application of modern automated customer counting tools may provide more accurate data about customer flows, but it cannot connect such 'anonymous' data with real customer properties captured by our survey. With the *Variable Clumping Method* (*VCM*), we introduce a new method to retail property research that allows a better analysis of concentration patterns with multiple scales. By introducing relative coupling to our analysis, we provide further insights into the important field of tenant placement and customer behaviour. Since our focus is on the spatial pattern of coupling, we do not adopt an explicitly behavioural perspective. This aggregated approach might neglect psychological aspects of individual behaviour, but could also contribute to a more profound view of actual coupling patterns and could enhance, for example, multi- agent simulations.

The spatial aspects, which are included in almost every investigation dealing with shopping centre tenant placement, suggest the possibility to use *GIS* from two perspectives: (i) analysing the data using spatial *GIS*-tools and (ii) generating data for spatially based models, such as regression models. Regarding previous results from shopping centre research on the topic 'tenant selection', the retail category concentration, pass ratio, coupling as well as the distance of shops to entrances, to the mall centre, to vacancies and in particular to other shops from the same or other retail categories, can be identified as the most important locational determinants of shopping centre rents. These factors determine the number of customers passing the shops, their shopping behaviour and thus ultimately the shops' turnover. The following aspects have been considered in this study and are analysed statistically, but also by spatial and visual analysis:

(1) Category concentration
(2) Customer flows
(3) Coupling

### 4.3.1 GIS model

Based on the catalogue of important shopping centre rent determinations above, the use of GIS as a shopping centre (research) management tool is presented in this section, using the *GIS ESRI ArcGIS* and the statistical programming language *R* (R Core Team, 2013). In this context, the GIS-application is exemplified in the context of a German shopping centre of about 60,000sqm *gross lettable area*, about 140 shop units, two storeys, about 35,000 customers daily and a floor plan as shown in Figure 1. Before presenting the possible uses of *GIS*, it is necessary to understand how a GIS-system works and which of its components are important. A GIS can be defined and described in terms of four structural and functional components (see Figure 2). In particular, the functional components - the so-called '*IMAP*s' - are of great practical importance. In this sense, a *GIS* is a computer-based system that enables the digital gathering (input), storage and management, analysis and presentation of spatial data (Ehlers and Schiewe, 2012). This is the functional structure on which the presentation of potential GIS-applications within the shopping centre management is based.

**Figure 1: Floor plan of the assessed shopping centre.**


**Figure 2: Four component model of a GIS.**

| **Four Component Model** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Structural Components** | | | | **Functional Components** | | | |
| Hardware | Software | Data | User | Input (**I**) | Management (**M**) | Analysis (**A**) | Presentation (**P**) |

*Source: De Lange et al. (2006, pp. 322ff.), Longley et al. (2005, p. 24), Segerer (2011, p. 28).*

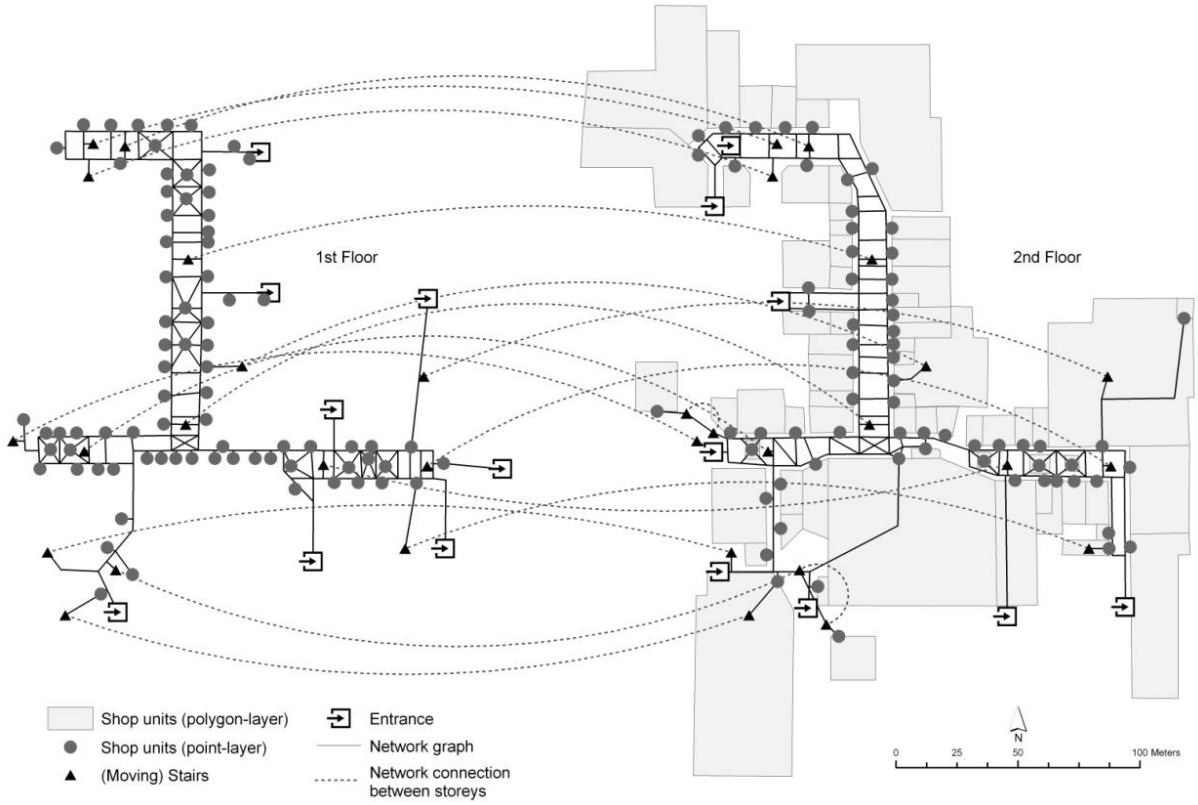
#### 4.3.1.1 Data input

The big advantage of GIS consists of connecting non-spatial database information with a matching geometry. Thus, the task of *GIS* as an input tool is to create a geometry of shop units and entrances

within the mall, as well as connections between these objects across different storeys. Therefore, a polyline-layer and a polygon-layer were created to represent the customer flow and the shop arrangement. The shop polygons were drawn according to a georeferenced plan of the shopping centre.

The big advantage of *GIS* consists of connecting non-spatial database information with a matching geometry. Thus, the task of *GIS* as an input tool is to create a geometry of shop units and entrances within the mall, as well as connections between these objects across different storeys. Therefore, a polyline-layer and a polygon-layer were created to represent the customer flow and the shop arrangement. The shop polygons were drawn according to a georeferenced plan of the shopping centre.

**Figure 3: Basic GIS network model.**




*Note: The polygon-layer of shops is not visualised in 1st floor here for reasons of clarity.*

Finally a point-layer was created, in which every shop unit was represented by its (single- point) entrance. This point-layer and the previously mentioned polyline-layer – with connections to the shop entrances – formed the basis for further spatial analysis. Therefore, we ended up with a *GIS* model – as a basis for data management and analysis – as shown in Figure 3. A customer survey was organised in which 1163 customer preferences, 1084 customer running routes and separately measured customer frequencies were observed. The survey was conducted at all entries / exits of the centre on four days (Tuesday, Wednesday, Friday and Saturday) in August, always between 10 am and 6 pm. The running routes were marked on a map (supported by a large plan of the centre with logos of all shops for better orientation) by the customers themselves with the help of the interviewers. Based on this survey, the answers were transferred to a database by storing the

attribute data as well as the geometry. The data from this customer survey was supplemented by information about the shop units (retail category, shop size and some additional information from a parallel survey of the shops' tenants; such as with questions about their estimation of the positive or negative effects of their neighbours' shops and their satisfaction with the shop location The results of this survey can be merged with the customer survey, but this is not the focus of this article).

#### 4.3.1.2 Data management

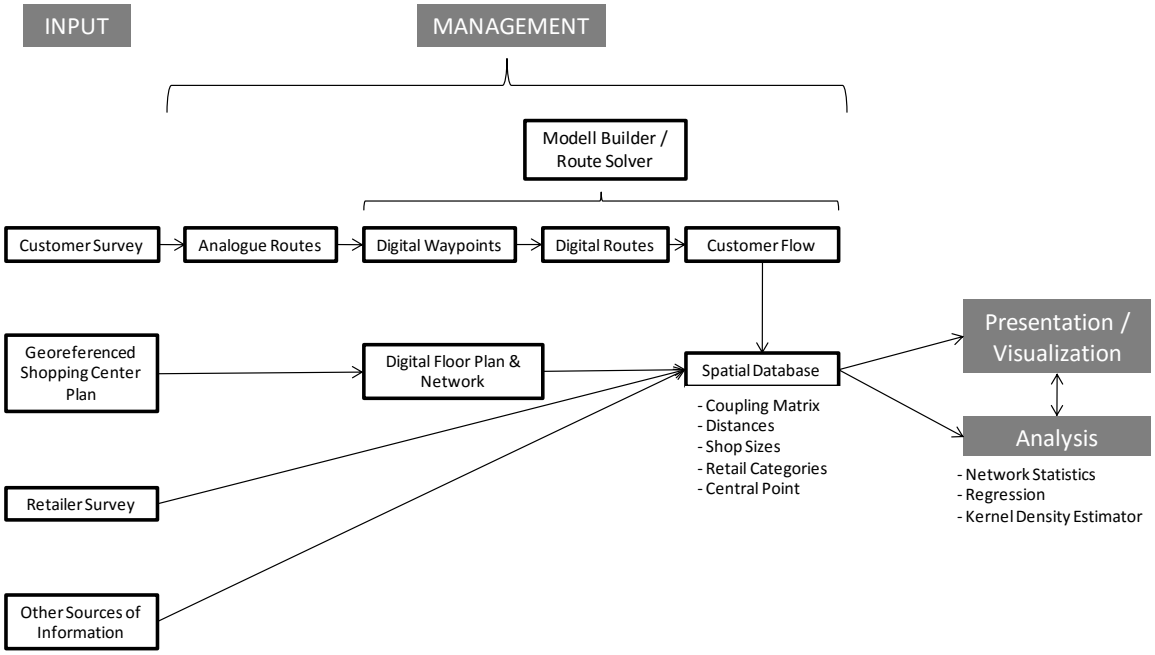
A GIS is a basic way of very effectively managing spatial data, like the walking paths of shoppers in a shopping centre. Before beginning the analysis, the data from the survey has to be prepared. In this context, the main task of *GIS* is to transfer the collected customer routes from the analogue survey − which is done with pen and paper − into a digital dataset. Therefore, the surveyed routes were divided into single points representing:

- the entrances and exits
- the shops and service providers visited
- further waypoints

The use of further waypoints in addition to the entrances and visited shops enables the representation of the entire route of one customer.

The application of *ArcGIS Model Builder* enables an automated calculation of 1084 customer routes, automating the route-solving process and further analysis. Additionally, the attribute data from the customer survey were connected to the route data. The connection of the geometric and the attribute data is the basis for analysing customer routes by spatial as well as by attribute characteristics (see Figure 4). Thus the built-up *GIS* combines the advantages of a 'classic' and of a spatial database, handling geometric and attribute data in one system. The database enables the comparison of the spatial behaviour of customers from different origins, customers visiting particular shops or male and female customers.

**Figure 4: Work-flow model in GIS to display the customer flow.**



### 4.3.2 Concentration of retail categories

#### 4.3.2.1 Category agglomeration

Thanks to its spatial character, the analysis of the shop arrangement is a typical field of application for a GIS. There are several ways of detecting a spatial clustering of retail categories within a shopping centre. When following the geographic retail research approach (see Heinritz et al., 2003) an agglomeration (or cluster) can be defined as a gathering of at least three retailers within a particular distance. The distance has to be derived from the dimensions of the specific problem and was defined as double the value of the mean length of shop windows, which is 16.5m (thus, the cluster distance is 33m). For the first floor, clusters in three retail categories – based on the *NACE* classification system (see Table 1) – could be observed (see Figure 5):

- food (two clusters)
- body & health
- fashion

It appears that the body and health cluster was located at the edge of the first floor, whereas the fashion cluster was located more centrally. Beyond that, the fashion cluster extended over a comparatively large area.

**Table 1: NACE retail category classification**

| | |
|---|---|
| **47.1** | Retail sale in non-specialised stores |
| **47.2** | Retail sale of food, beverages and tobacco in specialised stores |
| **47.3** | Retail sale of automotive fuel in specialised stores |
| **47.4** | Retail sale of information and communication equipment in specialised stores |
| **47.5** | Retail sale of other household equipment in specialised stores *(including textiles, electrical household appliances, furniture, lighting equipment and others)* |
| **47.6** | Retail sale of cultural and recreation goods in specialised stores *(including books, newspapers, music and video recordings, sporting equipment, games and toys and others)* |
| **47.7** | Retail sale of other goods in specialised stores *(including clothing, footwear and leather goods, cosmetic and toilet articles, flowers watches and jewellery and others)* |
| **53.1** | Postal activities under universal service obligation |
| **56.1** | Restaurants and mobile food service activities |
| **56.3** | Beverage serving activities |
| **64.1** | Monetary intermediation |
| **79.1** | Travel agency and tour operator activities |
| **79.9** | Other reservation service and related activities |
| **86.2** | Medical and dental practice activities |
| **93.1** | Sports activities |
| **95.2** | Repair of personal and household goods |
| **96.0** | Other personal service activities |

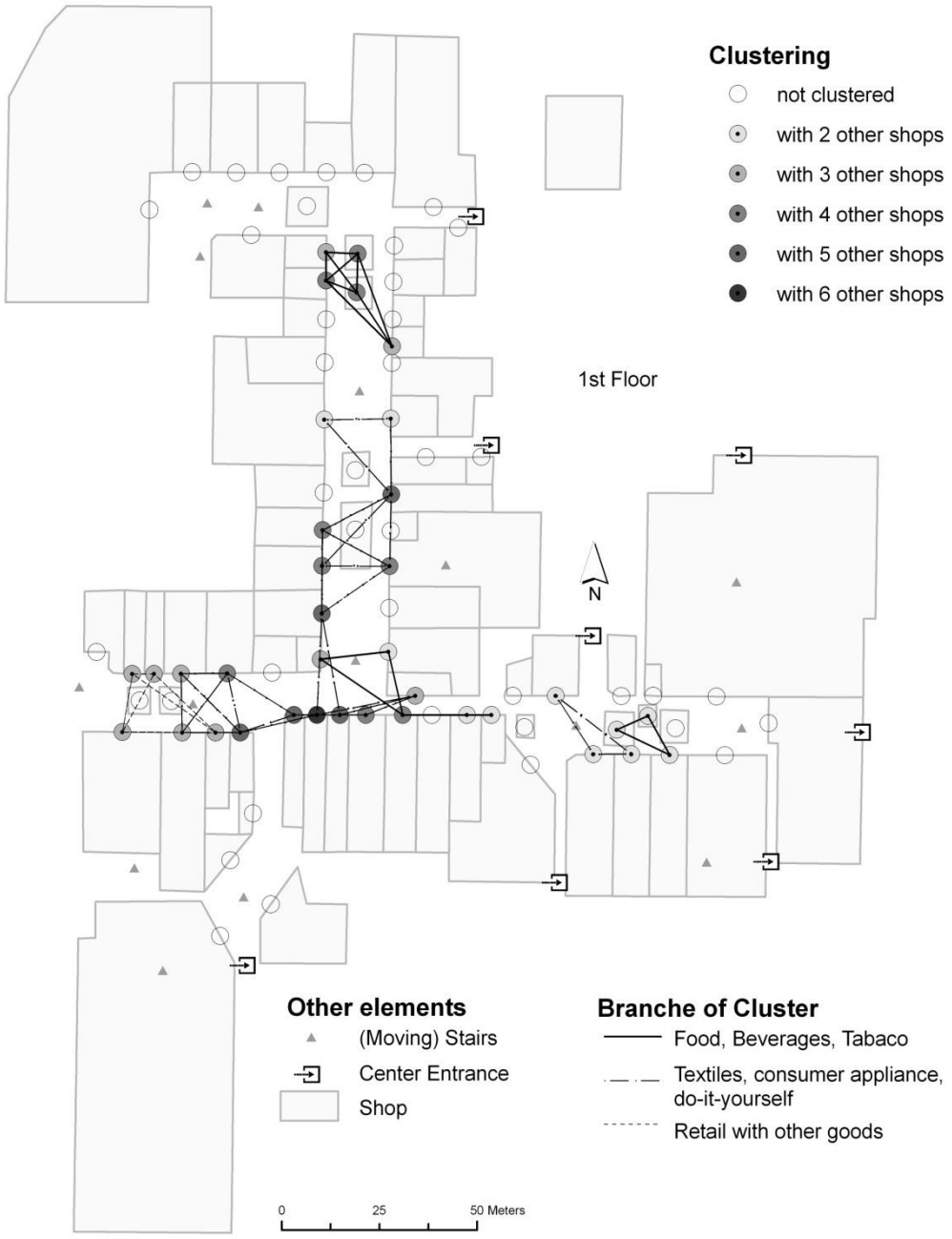
*Source: Eurostat (2008)*

#### 4.3.2.2 Variable clumping method

Another powerful and fairly similar way to search for clusters is to detect significant clumps in the distribution (Roach, 1968). Contrary to the geographic retail research approach, the clumping method does not require a minimum number of three units, but this restriction could be introduced

in the present context of shopping centre research. Okabe and Funamoto (2000) developed the concept of the *VCM* which allows for a multi-scale analysis that is able to detect concentrations on different spatial levels, as well as pinpointing of the actual shops constituting these clusters (see Figures 6 and 7). The multi-scale feature is very useful in our research context, because it also allows for the consideration of different mean shop sizes in different retail categories. *VCM* could be also used to analyse the distribution of shop sizes within the shopping centre, by using the classified sales area instead of a shop's retail category. The results of *VCM* on category concentration can be used as input variables for further research, which analyses the determinants of shopping centre rents and sales (Des Rosiers et al., 2009) and can consequently improve the identification of 'ideal tenant placement'.

**Figure 5: Clustered retail branches**

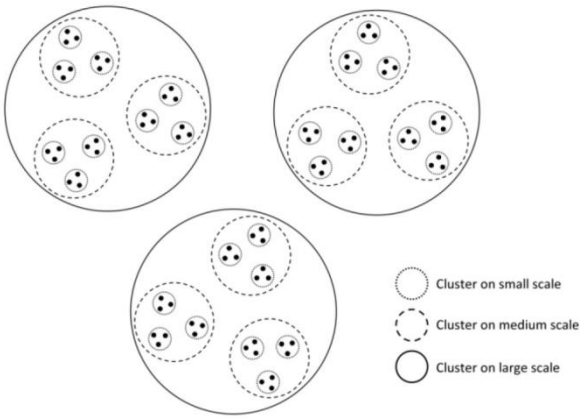


The challenge of applying this method to the present shopping centre was the irregular bounded area to be analysed. Existing studies, that have used the clumping method, conducted an analysis of unbounded or regular bounded areas or networks.

Many common methods for the determination of retail category concentration like the *Herfindahl index* measure only the proportion of sales area of different retail categories and do not detect actual spatial clusters. In many cases, this can yield misleading results. The *Herfindahl index* produces equal results for the two substantially disparate cases of one cluster of eight shops and two clusters of four shops in two separated regions of a shopping centre. *VCM* has the ability to differentiate between these differing cases and can furthermore consider the significance of each cluster by applying a probabilistic approach. The following quotation from Okabe and Funamoto (2000, p. 112) gives a short description of the *VCM* and introduces the relevant terms of clump, clump radius and clump size:

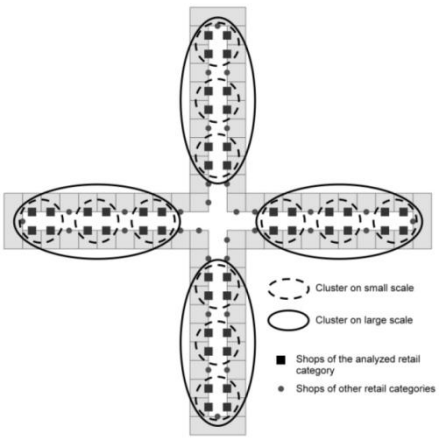
> *"It is a class of methods for finding 'clumps' in the distribution of points, $p_1$; … ; $p_n$, over a bounded region S. Usually a clump is defined in terms of circles centred at given points $p_1$; … ; $p_n$ […]. The radius of the circles is called a clump radius. A clump is then defined as a set of points whose circles are connected. The number of connected circles in a clump is called the size of the clump."*

**Figure 6: Idealised setting of multi level clumps**




*Source: Adaption from Okabe and Funamoto (2000, p. 112).*

**Figure 7: Multi level clumps within a shopping centre**



The so-called clump state depicts the encountered number of clumps for all possible clump sizes, and a range of clump radii. According to *VCM*, a clump can be referred to as significant, if the empirically observed number of clumps of size *i* for a given radius *r* is significantly larger than the number of clumps that would appear in the distribution of random points. By applying *Monte Carlo simulation*, a critical number of clumps is determined for different levels of significance and finally compared with the empirical number of clumps. Assuming a very high number of shops from one retail category, *VCM* would denote the resulting clump(s) as not significant, because it is quite natural for these shops to be close together. Likewise empirical clumps might be not significant, if a very high clump radius is considered.

**Table 2: Number of clumps for retail category 47.5**

| Number of Clumps | *Clump size i* | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Clump radius rj [m]* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* | *17* | *18* | *19* | *20* | *21* |
| *5* | 13 | 2 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *10* | 2 | 0 | 1 | 0 | 0 | 0 | **1** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *15* | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 |
| *20* | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| *25* | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| *30* | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| *35* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *40* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *45* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *50* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

not significant
**significant at p<0.10**
**significant at p<0.05**
**significant at p<0.01**

The clump state that results from a random distribution of shops is obtained by conducting a *Monte Carlo simulation*. Unlike Okabe and Funamoto (2000) who distribute the random points on a plane, and Shiode and Shiode (2009), who developed a network-based version of *VCM*, we distribute the shops along the edges of our shopping centre and assume that customers can move freely within the mall without any restriction to a given network. A simple construction of the clumps by generating circles around the investigated points is not possible in this case. Okabe and Funamoto (2000) suggest the use of *Minimum Spanning Trees* (*MST*) instead. The *MST* is the shortest subgraph connecting all vertices – in our case, shops of one retail category – of a given undirected graph. The *MST* enables an easy analysis of existing clumps in the graph, but it has to be generated anew for each *Monte Carlo* step and the empirical distribution. Okabe and Funamoto (2000) apply Voronoi diagrams (for their rectangularly bounded plane) and Delaunay triangulation for obtaining the minimum spanning tree. Due to default of a standard solution in common *GIS* software for the generation of *Voronoi diagrams* in irregularly bounded planes forming a non-convex set (like a shopping centre), we apply the algorithm of Prim (1957) to determine *MST*, beginning with a distance matrix of all shops from one retail category. The distance matrix is generated using the algorithm of Dijkstra (1959), beginning with a network of connections between all shops and all concave vertices of the shopping centre's edge. The computation of VMC is done entirely with the statistical software R (R Core Team, 2013). The procedure is standardised, and arbitrary shopping centre plan configurations and distributions of shops can be used as input via simple ASCII files.

The following interpretation of *VCM* results focuses on the retail category 47.5 'Retail sale of other household equipment in specialised stores' which includes the sale of textiles, electrical household appliances, furniture, lighting equipment and others. Table 2 shows the number of significant and non-significant clumps of different sizes for clump radii from 5 to 50m, in steps of 5m. The table shows, for instance, a significant clump of four shops for a clump radius of 5m on a significance-level

of $p < .1$ ($n_{Monte\ Carlo}$ = 1.000). In the case of a clump radius of 35m, all shops form one clump which, however, is not significant.

Figure 8 gives an overview of the location and significance of clumps for different clump radii and shows the structure of the *MST*. Table 2 and Figure 9 reveal a clustering of category- 47.5-shops with a medium spatial scale in the southern half of the analysed storey and a clustering with a small scale, south and east of the middle of the centre (*MC*). Table 3 displays an overview of significant clumps that could be identified for all retail categories. In accordance with the assumptions of Yuo and Lizieri (2013), the concentration of retail categories is rather low (except for the category 47.5), in our case of a two-storey shopping centre with a moderately complex floor plan (the complexity value in terms of Yuo and Lizieri (2013) is 0.033). In terms of optimising the tenant placement, the existing situation can be described as very good.

**Figure 8: Number of clumps for retail category 47.5**

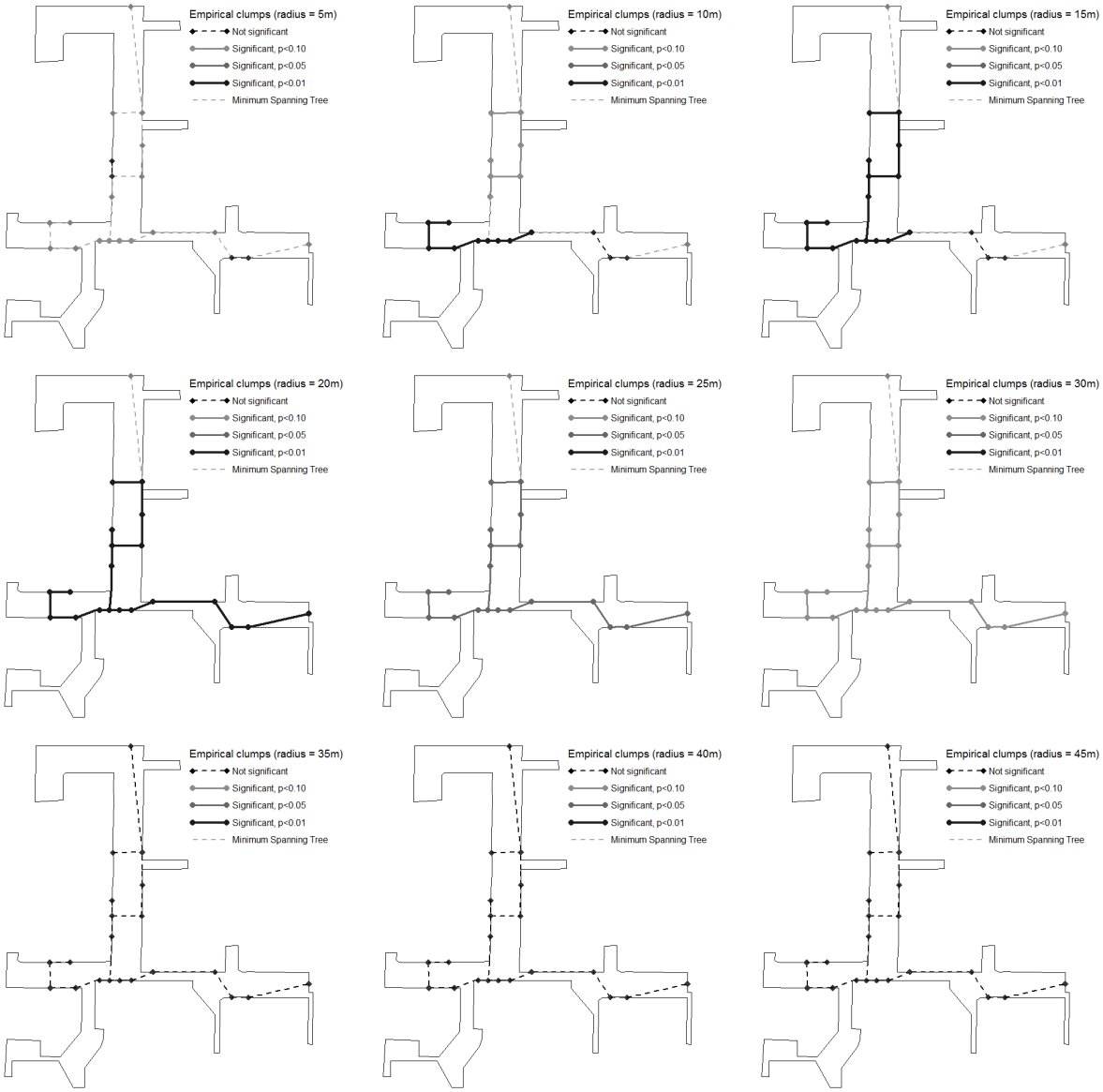


Figure 10 condenses the results of the *Monte Carlo simulation* ($n_{Monte\ Carlo}$ = 2.500) for retail category 47.5 and continuous clump radii from 1 to 50m in steps of 1m. This demonstrates the decreasing (increasing) share of shops that are part of clumps with small (large) size with a growing clump radius and the significance of empirical clumps in relation to their probability of occurrence in the randomly simulated distribution.

**Figure 9: Share of shops constituting clumps of a certain size for continuous clump radii from 1 to 50m in steps of1 m according to Monte Carlo simulation (circles: empirical clumps with number of shops and level of significance).**

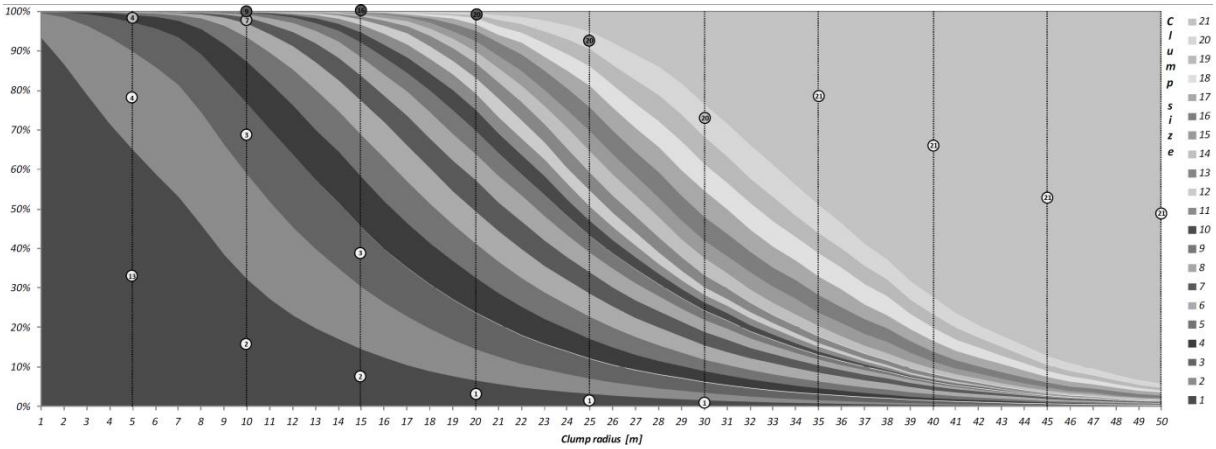



**Figure 10: Passers-by ratio, buy ratio, pass-buy ratio**

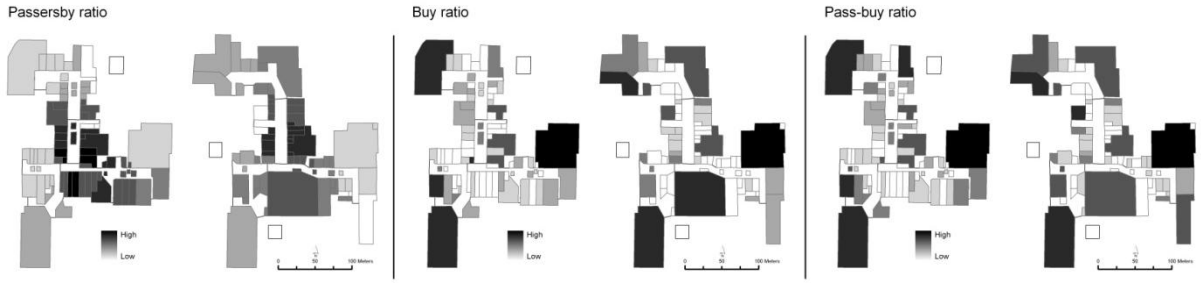
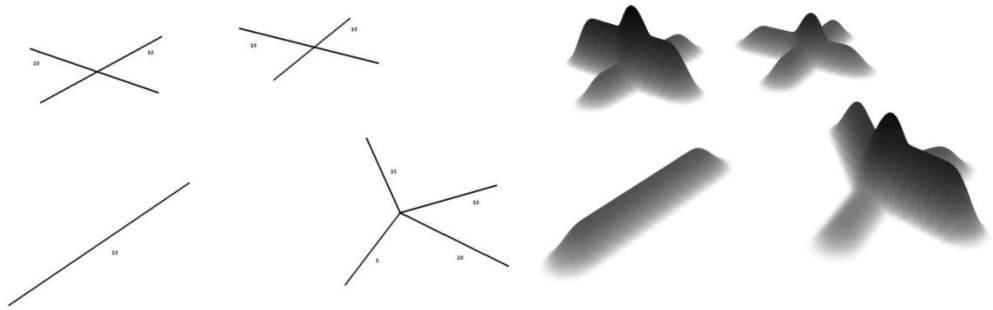



**Table 3: Significant clumps for all retail categories**

| ***Retail category*** | ***Number of shops*** | ***Significant clumps*** *("number of clumps" x "clump size" @ "clump radius"; ***: p<0.01, **: p<0.05, *:p<0.1)* |
|---|---|---|
| *47.1* | *3* | no significant clumps |
| *47.2* | *13* | **1x6**@10m**, **1x6**@15m*, **1x8**@35m* |
| *47.4* | *2* | no significant clumps |
| *47.5* | *21* | **1x4**@5m* , **1x7**@10m*, **1x9**@10m*** , **1x16**@15m*** , **1x20**@20m*** , **1x20**@25m** , **1x20**@30m* |
| *47.6* | *6* | **2x2**@10m*, **2x3**@25m**, **2x3**@30m**, **2x3**@35m**, **2*3**@40m*, **2x3**@45m**, **2x3**@50m** |
| *47.7* | *11* | **1x11**@30m** |
| *53.1* | *1* | --- |
| *56.1* | *2* | no significant clumps |
| *64.1* | *1* | --- |
| *79.1* | *3* | 1x2@5m** |
| *95.2* | *1* | --- |
| *96.0* | *2* | no significant clumps |

**Figure 11: Kernel density estimators for (weighted) line features**



#### 4.3.2.3 Analysis of distances between shops from same or different retail category

As well as using the *VCM*, it is also possible to analyse the mean distances between the shops, in order to detect signs of retail category clustering. The calculated mean distance between one shop and all other shops is 127.6m. Variations of this value between the different retail categories depend on their respective location within the centre (lower values in the 'middle' of the centre). Furthermore, there were variations in value between the shops of similar and of other retail categories, which suggests the existence of clusters (see Table 4).

Using a *student's t-test*, a significant lower mean distance between shops of the same retail category (117.6m) than between the shops of different categories (129.2m) can be assessed. Fashion stores have the lowest mean distance between each other (108.2m). This value is already low, because of the central location of the fashion stores, but the mean distance to shops of other retail categories is significantly higher (121.1m). The comparison of the values in Table 4 shows that the retailers for domestic appliance, fashion/textiles & DIY and those of foodstuffs, drinks & tobacco tend to be clustered, because retailers of the same category are located significantly closer together than those of other categories. In contrast, retailers for published goods, sports equipment & toys, restaurants, snack bars, cafés & ice cream parlours, communication devices and information, and retailers with goods of different types are arranged further apart.

### 4.3.3 Customer flows

Besides the spatial concentration of retail categories, the use of the basic network model, combined with customer survey data, makes it possible to identify areas of high or low ('dead spots') customer attention. This is operationalised through the passers-by ratio (*PR*) reflecting the proportion of all customers that pass a certain shop or area of the *SC* within a certain period.

Figure 10 shows the calculated *PR* which could, for example additionally be compared to the retailers' self-assessment of their site quality (which was ascertained in a simultaneous survey). Another possibility for describing customer flows within a shopping centre includes establishing the proportion of all customers who enter particular parts of the centre. This passers-by ratio had already been calculated by Brown (1992) but with the help of a *GIS*, the work can be done much faster and in greater detail. In this case, the shopping centre was partitioned into 53 mall sections. In fact, about half of all customers walk through the most centrally located part of the centre. These spatial data are thereupon combined with the data about actual shopping behaviour.

For the operators of a shopping centre, it is necessary to know which shops are able to transform a high PR into high sales and which are not. Brown (1992) introduced the concept of a pass-buy ratio (PBR) to embody this fact in research. The *PBR* is calculated from the buy ratio (*BR*) and the *PR* (Equation 1). The *BR* indicates which fraction of all customers buy something in one particular shop. If, for instance, every second customer (*PR* = 0.5) passes a particular shop and every fourth customer (*BR* = 0.25) buys something in this shop, the *PBR* will be 0.5, meaning that every second customer passing the shop purchases something. The values of *PBR*, *BR* and *PR* are displayed in Figure 10. Especially some fashion stores in the first floor have a very low *PBR*, indicating some optimisation potential, because these shops are not able to generate revenue from the high *PR*s in this location.

$$PBR = {}^{BR}/_{PR} \qquad (1)$$

where *PBR = pass-buy ratio, BR = buy ratio, PR = passers-by ratio.*

Another useful method for visualising customer frequencies is the application of *kernel density estimators* (*KDE*). *KDE* is a non-parametric method of estimating an unknown distribution, for example, that of customers in *SC*. In fact, density estimation along the edges of our network can be difficult to calculate correctly and does not correspond to the rather laminar nature of people walking through a shopping centre. Typically, *KDE* are used to generate density values based on point data that are arranged on a surface or along a network. For instance, *KDE* can be used to analyse the frequency of natural hazards in a particular region, the risk of traffic accidents in an urban street network (Okabe et al., 2009; Xiea and Yanb, 2008) or the frequency of crimes in a city (Nakaya and Yano, 2010). In the present case, *KDE* results in a curved surface, representing the probability of finding a customer in a certain area of the shopping centre (see Figure 11).

As is customary in *KDE*, the surfaces fitted over all single lines i are added up to obtain the overall density surface respective to the estimated density at point *x* (Equation 2):

$$\hat{f}_{k,H}(x) = \frac{1}{nh}\sum_{i=1}^{n} K_i(x) \qquad (2)$$

where $\hat{f}(x)$*:* estimated *kernel density* at location *x, K:* kernel function*, n:* sample size*, h: bandwidth.*

The volume under this surface represents a probability and adds up to 100%. The shape of the surface is determined by the chosen kernel function (Equation 3) and bandwidth. We use a bounded quartic kernel function, which is a modified version of the *normal biweight kernel*. 'Bounded' means that those areas outside the bandwidth do not contribute to the overall density.

The kernel function can be represented as:

$$K_i(x) = 3\pi^{-1}\left(1 - \left(\frac{x-x_i}{h}\right)^2\right)^2 \quad \text{if } \left(\frac{x-x_i}{h}\right)^2 < 1$$
$$K_i(x) = 0 \quad \text{otherwise} \qquad (3)$$

where $K_i(x)$*: kernel density* at location *x, xi:* location o*f* $i^{th}$ observation*, h: bandwidth.*

**Table 4: Calculated shop-to-shop distances**

| *Retail Category* | *Description* | *Same category* | *n-same* | *Different Category* | *n-different* | *All shops* | *p-Value* |
|---|---|---|---|---|---|---|---|
| ***47.5**** | ***Domestic appliance, fashion/textiles, DIY, accommodation*** | 108.2 | 36 | 121.1 | 105 | 117.9 | 0.000 |
| ***47.2**** | ***Foodstuffs, drinks, tobacco*** | 118.6 | 29 | 123.9 | 112 | 122.8 | 0.030 |
| ***47.7*** | ***Other goods*** | 119.2 | 19 | 124.5 | 122 | 123.8 | 0.127 |
| ***47.6*** | ***Published goods, sport equipment, toys*** | 138.8 | 15 | 132.6 | 126 | 133.2 | 0.192 |
| ***56.1*** | ***Restaurants, snack bars, cafés, ice cream parlours*** | 130.2 | 11 | 124.5 | 130 | 124.9 | 0.336 |
| ***96.0*** | ***Other (personal) services*** | 157.3 | 7 | 138.8 | 134 | 139.6 | 0.107 |
| ***47.4*** | ***Devices of communication and information technology*** | 163.1 | 6 | 143.4 | 135 | 144.1 | 0.107 |
| ***79.1*** | ***Travelling agencies*** | 132.8 | 4 | 136.2 | 137 | 136.2 | 0.862 |
| ***64.1*** | ***Banks*** | 148.0 | 4 | 142.7 | 137 | 142.9 | 0.786 |
| ***47.1*** | ***Retailers with goods of different types*** | 197.2 | 4 | 173.7 | 137 | 174.2 | 0.277 |
| ***53.1*** | ***Postal services*** | | 1 | 112.9 | 140 | 112.9 | |
| ***56.3*** | ***Selling beverages*** | | 1 | 113.9 | 140 | 113.9 | |
| ***79.9*** | ***Reservation services*** | | 1 | 121.8 | 140 | 121.8 | |
| ***95.2*** | ***Repairing of consumer goods*** | | 1 | 130.6 | 140 | 130.6 | |
| ***93.1*** | ***Fitness centre*** | | 1 | 156.4 | 140 | 156.4 | |
| ***86.2*** | ***Doctor's and dentist's surgeries*** | | 1 | 163.2 | 140 | 163.2 | |
| ***Mean**** | | 117.6 | | 129.2 | | 127.5 | 0.000 |

****:* Significant on a 1%-level ($p < 0.01$)
***:* Significant on a 5%-level ($p < 0.05$)

The final result of a *KDE* depends more on the chosen bandwidth than on the actual kernel function. Additionally, the statistical efficiency of the different kernels differs only slightly (Silverman, 1986, p. 43). The rules for choosing the optimal bandwidth specified in the literature 'are not directly applicable to density estimation on networks' (Okabe et al., 2009, p. 24). The value has to be chosen by a basic consideration of the nature of the problem. For example, Okabe et al. (2009) chose a bandwidth of 200m in their estimations of traffic accident densities. We used a bandwidth of 15 m to represent the typical proportions of a shopping centre. A shopping centre with other proportions, as well as another network configuration, can make another bandwidth more practicable. Additionally, we weighted the routes according to the centre's entrance where they had been collected and its typical customer frequency, which was ascertained by the centre management before our survey was conducted.

**Figure 12: Identifying 'dead spots' by kernel density estimation**


**Figure 13: Chronological sequence of customer flow during the course of one week**

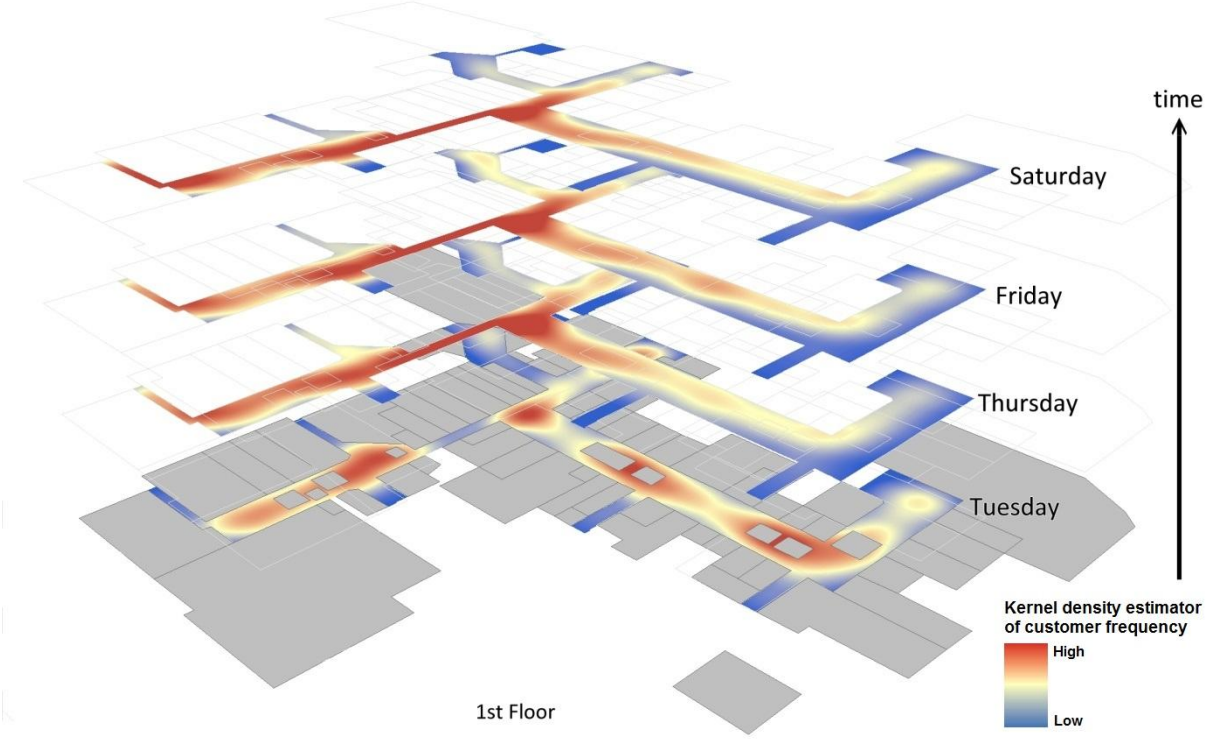


**Figure 14: Identifying the mall-centre (MC)**


**Figure 15: Visual analyses of the passers-by ratio**

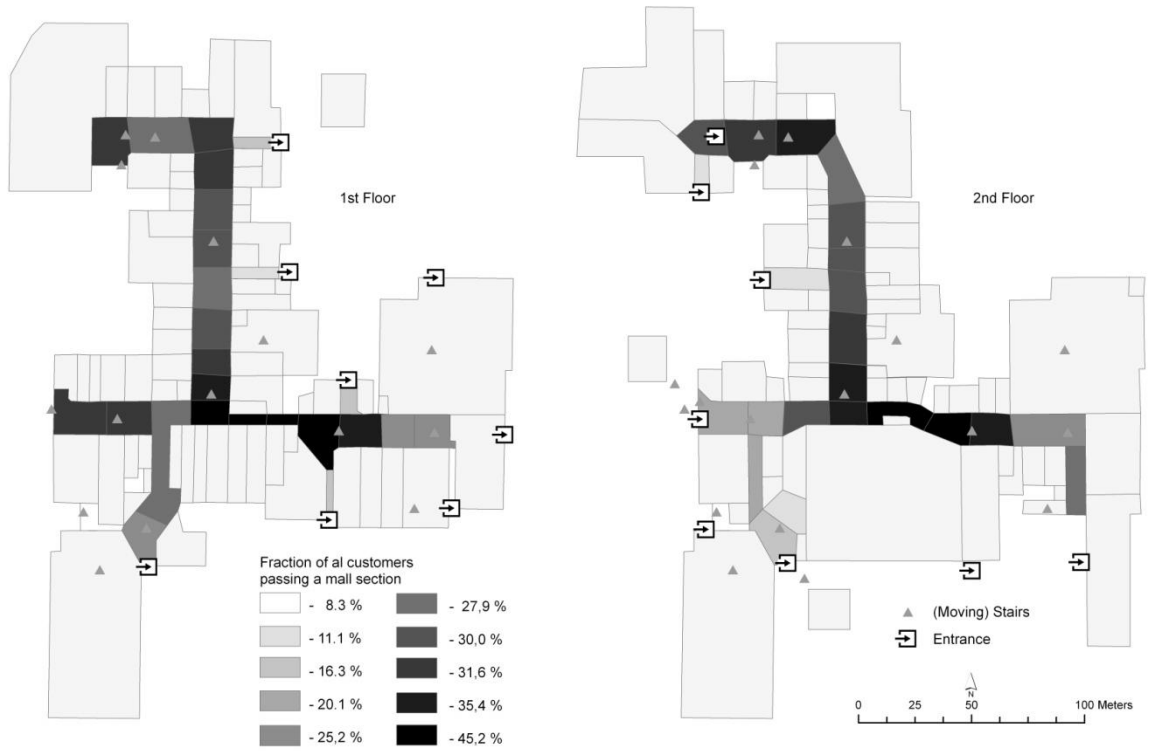


**Figure 16: Correlation between the distance to mall-centre and the passers-by ratio**

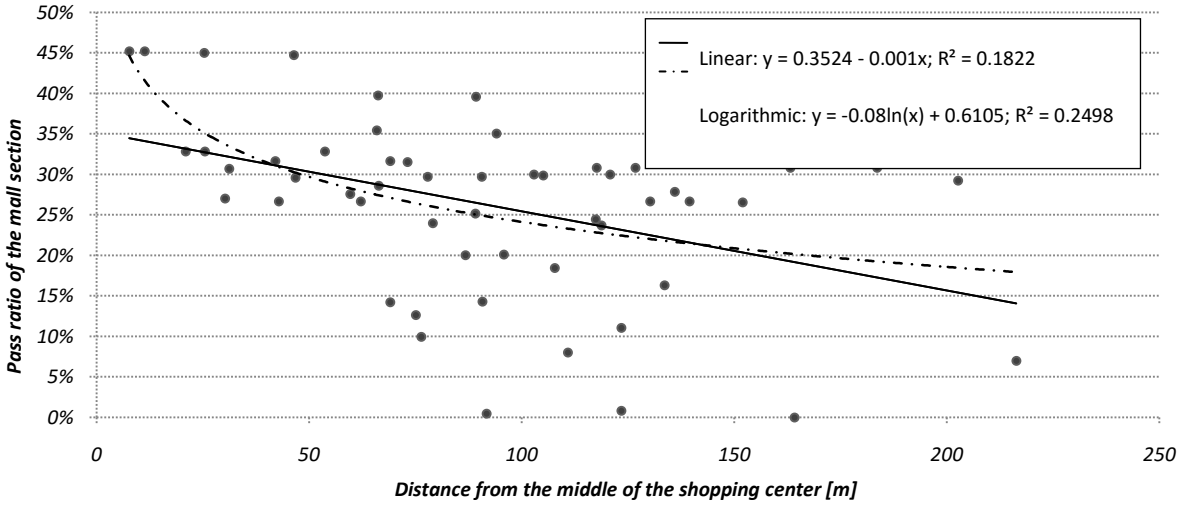



Figure 12 shows the result of the kernel density estimation. There is an evidently higher density of customer flows in the centre's second floor. The first floor shows the expected pattern of distinctly higher density value in the middle of the floor, whereas the second floor shows a significant exception to this pattern in the northern aisle. Figure 12 also indicates some potential for optimisation in the eastern parts of both floors, such as by locating more attractive shops there.

Because the conditions within a shopping centre might change during the course of one year or even in one day, the temporal aspects of customer flow are a matter of concern, especially for the operator of the centre. With regard to analysing punctual information (e.g. crimes) Nakaya and Yano (2010) developed a method for calculating a 3D kernel density which integrates space as well as time to analyse a chronological sequence of events. In the case of an analysis of linear data, such a procedure does not yet exist. In any event, *KDE* can be calculated separately for different time frames, for example the four single days of our survey (see Figure 13), but we could also aggregate customer flows of the same time frame for different days.

The *KDE* can be used to analyse the dependency of the passers-by-frequencies and the location within the centre. Previous studies state unanimously that the frequency decreases with a rising distance from the central point of the shopping centre. This point is known as the *MC*. First of all, in order to test this hypothesis, it is necessary to locate a point which represents the *MC*. In a shopping centre which has a complex floor plan like the one we analysed, it is not easy to find a point which can rightly be called the 'centre' or the 'middle'.

One way to define the *MC* is to find the point of maximum distance to the next entrance. This method has the flaw that it can only take into consideration one entrance per point and can thus generate misleading results. A better alternative for ascertaining the *MC* is to calculate, for every point in the mall, the mean distance to all entrances. This method obviously represents the accessibility of a point within the centre. Therefore, the *MC* calculated by this method is used for all further analysis with reference to the central point. In our case, the point is located in a very well frequented spot on the first floor (see Figure 14). A more theoretical approach would be to calculate the median point, namely that with the least distance to all other points. Admittedly, this point does not consider the number and location of entrances of the shopping centre and consequently does

not represent actual accessibility. In any event, it turned out to be located quite close to the point of minimum distance to all entrances. To check the hypothesis that the frequency decreases with a larger distance from the centre, a point grid can be used to ascertain the value of the *KDE* for a particular location. The (network) distance of these points to the *MC* can be used as an independent variable in a regression analysis of the passers-by-frequency. In fact, the highest values of passers-by-frequency appear at the *MC* and decrease with a rising distance from the *MC*. Another passers-by-hotspot is located exactly at the maximum distance from the *MC* near a magnet store. However, the regression model – from a statistical perspective – rejects the hypothesis of a linear correlation between the distance to the *MC* and a decreasing frequency using *KDE* (see Figure 16). The residuals of this regression were mapped and scanned for spatial autocorrelation. They show a significantly clustered pattern of distribution (*Moran's I*), which continues along with the high frequency values near magnet stores very distant from the *MC*. Future research could consider further variables in this regression.

A more significant coherence between the distance from the *MC* and the customer frequencies can be found by taking into consideration the pass ratios of particular mall sections (see Figure 15). There is an obvious decline in the pass ratios with longer distances from the *MC*, where the highest ratios can be observed. We found the best fit for a logarithmic relationship. A scatterplot of the *PR*s and their corresponding distances from the *MC* plus a regression, can illustrate this observation (see Figure 16).

### 4.3.4 Coupling

Another interesting aspect of customer behaviour in a shopping centre is the coupling of multiple shop visits in one single shopping trip. The coupling depends inter alia on the retail category (Nelson, 1958), and understanding it provides a fundamental way for shopping centre operators to control the flow of visitors within a building. In this case, all the shops where the customers spend money were surveyed and the results merged in a 'coupling matrix'. Generally, the coupling matrix provides information about which shops yield the highest coupling (e.g. the department store) and which shops are visited specifically without any further shopping (e.g. service providers like a laundry). This attribute of a shop is denoted as 'absolute coupling', which is represented by the column sum (equalling the row sum in this case) in the coupling matrix for particular shop.

It is also possible to diagnose the coupling between one particular shop and all the other shops. The relative coupling of a shop was calculated by using a method derived from global city research (Hoyler, 2005; Taylor and Walker, 2004). In order to calculate this relative coupling, for each shop *i*, we regress (Equation 4) its absolute coupling with other shops $C_{ij}$ against their summed absolute coupling with all other shops $C_{+j}$.

$$\hat{C}_{ij} = \beta_{0i} + \beta_i \cdot C_{+j} \qquad (4)$$

$\hat{C}_{ij}$ = estimated coupling between shops $i$ and $j$
$\beta_{0i}$ = intercept of shop $i$
$\beta_i$ = regression coefficient of shop $i$
$C_{+j}$ = summed absolute coupling of shop $j$
and $i \neq j$

The residuals of this regression provide information about the relative coupling behaviour.

$$\varepsilon_{ij} = C_{ij} - \hat{C}_{ij} \qquad (5)$$

$\hat{C}_{ij}$ = estimated coupling between shops $i$ and $j$
$C_{ij}$ = observed coupling between shops $i$ and $j$
$\varepsilon_{ij}$ = Residual value
and $i \neq j$

Positive residuals ($\varepsilon_{ij}$, Equation (5)) represent a higher amount of coupling than would be expected using absolute values. Figure 17 shows these relative coupling values for the supermarket within the investigated centre. As Taylor and Walker's (2004) approach indicates, each shop was represented by a simple square to avoid creating a visual bias by referring to different shop sizes. In the case of the supermarket, there were high values of coupling to the drugstores and bakeries, whereas coupling to fashion shops was rather low. One approach for maximising synergetic effects would be to increase the distance between these shops, because the large number of customers visiting the supermarket and one of the other shops would have to pass more other shops on their way. Similar results were observed for the organic food store, which shows high coupling with the supermarket and drugstores shop, whereas coupling with fashion stores was rather low. The department store showed high values of coupling to other fashion shops (except for young fashion), but only low levels of coupling with food shops. This coupling behaviour can also be seen in terms of the customers' routes within the shopping centre. As mentioned above, the use of *GIS* makes it possible to select particular routes according to the answers in the interviews – e.g. concerning visits to particular shops. These selected routes can subsequently be used to calculate *KDE*, which shows the mean probability of finding a visitor to one shop in a particular area within the shopping centre (see Figure 18). This figure suggests the placement of a shop with high coupling with the supermarket (e.g. drugstore) in the eastern part of the mall, in order to increase the presently low customer flow there.

**Figure 17: Visual analysis of relative coupling**

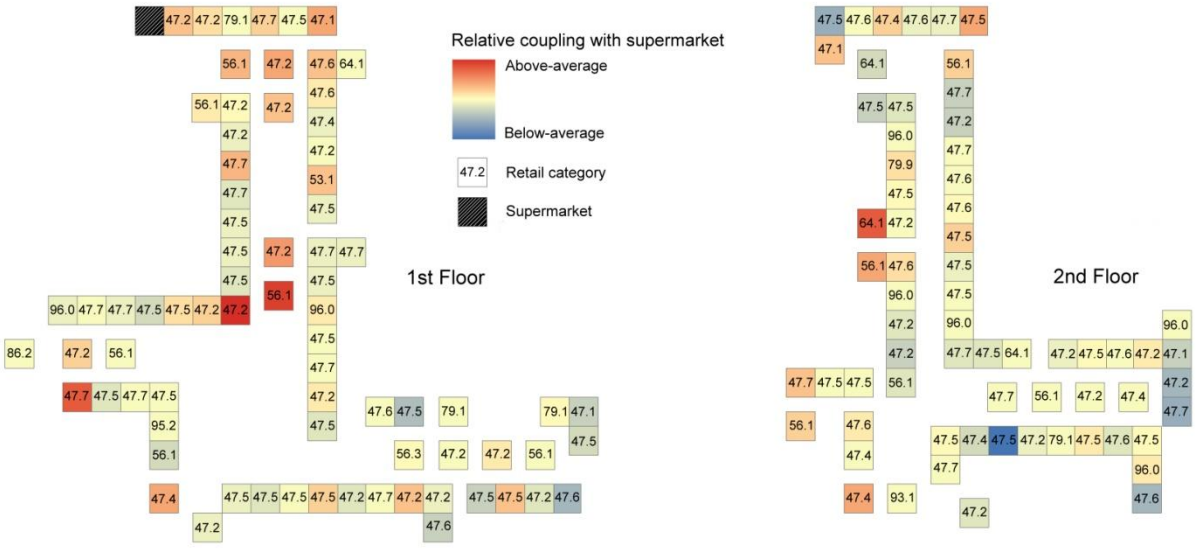


**Figure 18: Spatial movement of supermarket customers**


The *KDE* analysis demonstrates which routes the customers of one shop took (i.e. which entrance they used) and which other shops they passed during their shopping in the centre. The latter is of particular interest for operators of the shopping centre, because they are interested in maximising the amount of shop-to-shop coupling in order to maximise revenue. The customer flows – calculated by *KDE* and shown in Figure 18 – correspond explicitly with the relative coupling of the supermarket. There are obviously strong connections (flows) that can be identified within the centre as a whole. Other shops show a more spatially concentrated pattern. Table 5: Comparison of coupling between stores of the same or of another retail category

| ***Retail Category*** | ***All retail categories*** | ***47.1*** | ***47.2*** | ***47.4*** | ***47.5*** | ***47.6*** | ***47.7*** | ***53.1*** | ***56.1*** | ***56.3*** | ***64.1*** | ***79.1*** | ***79.9*** | ***86.2*** | ***93.1*** | ***95.2*** | ***96.0*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Other*** | 5.0 | 27.7 | 3.9 | 6.2 | 5.7 | 4.8 | 4.0 | 8.3 | 3.8 | 2.1 | 5.2 | 0.7 | 1.4 | 3.3 | 0.0 | 0.6 | 0.5 |
| ***Same*** | 7.2 | 159.7 | 2.9 | 11.9 | 10.9 | 6.3 | 2.4 | --- | 2.0 | --- | 3.7 | 0.0 | --- | --- | --- | --- | 0.0 |
| ***n*** | 141 | 4 | 29 | 6 | 36 | 15 | 19 | 1 | 11 | 1 | 4 | 4 | 1 | 1 | 1 | 1 | 7 |
| ***Sign.*** | 0.00 | 0.01 | 0.06 | 0.21 | 0.00 | 0.18 | 0.07 | | 0.07 | | 0.74 | 0.56 | | | | | 0.19 |

Significantly *(p < 0.05)* higher coupling with shops of the same retail category

Significantly *(p < 0.05)* higher coupling with shops of another retail category

Higher coupling with shops of another retail category

No values for 'same retail category'

**Table 6: Compatibility matrices showing coupling of different types of shops**

| ***All Retailers*** | *Similar shop type* | *Dissimilar shop type* |
|---|---|---|
| *Spatially proximate* | 132.0 | 83.7 |
| *Spatially separate* | 143.2 | 105.2 |

| ***Retailers with other goods*** | *Similar shop type* | *Dissimilar shop type* |
|---|---|---|
| *Spatially proximate* | 30.8 | 66.2 |
| *Spatially separate* | 66.7 | 84.9 |

| ***Food*** | *Similar shop type* | *Dissimilar shop type* |
|---|---|---|
| *Spatially proximate* | 69.1 | 75.6 |
| *Spatially separate* | 36.5 | 71.4 |

| ***Fashion*** | *Similar shop type* | *Dissimilar shop type* |
|---|---|---|
| *Spatially proximate* | 198.4 | *94.4* |
| *Spatially separate* | 217.8 | *122.5* |

It is not only interesting to consider the coupling between one shop and all other shops, but also of all shops in certain retail categories. Table 5 shows the calculated results of coupling values, differentiated for shops in the same retail category. The coupling of shops in the same category is significantly higher than those of other categories for retailers with varying types of goods (47.1), retailers with domestic appliances, fashion, DIY and furnishings (47.5) and the mean of all shops. In contrast, the coupling is lower between the retailers of foodstuffs, drinks and tobacco (47.2) than the coupling with other retailers.

Using *GIS*, it is not only possible to analyse absolute and relative shop-to-shop coupling, but also to analyse coupling with regard to distance. Brown (1992) demonstrated that coupling is higher if shops are from the same retail category and if they are located close to each other. Based on the calculated mean shop-to-shop distance of 127.6m, we split up shops into spatially proximate ones (distance between two shops ≤ 127.6m) and spatially separated shops (distance between two shops >127.6m). Using the distinction between shops of the same type and different retail categories, we obtained a 2×2-matrix (see Table 6). Contrary to Brown (1992), we find a higher coupling for spatially proximate shops only in the case of food, whereas coupling is higher for spatially separate shops in the case of fashion and when all shops are analysed, regardless of the retail category.

The assumption that stores with high coupling values are located at the sites of high customer frequencies could not be verified. Customer frequency at a particular store was operationalised through the *PR* of the mall section it lies within.

## 4.4 Conclusions and discussion

The previous demonstrations have shown that *GIS* improves and simplifies spatially based shopping centre analysis. At this point, it is necessary to discuss the results concerning the fields of research outlined in Section 4.3 and regarding the findings in the existing literature.

### 4.4.1 Category concentration

As Yuo (2010) demonstrated *GIS* is the ideal tool for analysing shop-to-shop distances. In terms of category concentration, *GIS* can provide the basis for identifying retail category 'clumps' and assessing retail category concentration using a mean comparison test (*student's t-test*).

Using a distance-based 'clump' analysis, clusters of three *NACE*-retail categories – food (two clusters), body and health and fashion can be identified. Obviously, concerning the retail categories and the methodology, our approach is very different to that of Carter and Haloupek (2002) (a *p-median model* vs. *VCM*). *VCM* enables a multi-scale analysis of category concentration. The results indicate a strong concentration of fashion stores and to a lesser extent of food retailers. The results of the *student's t-test*, which compared the distances between the different and similar retail categories, confirmed a concentration of the non-anchor stores in the categories of food and fashion. Thus, the overall distance- based analysis of the shop arrangement suggests – in contrast to the findings of Carter and Haloupek (2002) – a relatively low concentration of retail categories. According to Yuo and Lizieri (2013), this can be explained by the small number of storeys in our centre and its relatively low flow plan complexity.

### 4.4.2 Customer flows

The *KDE* is an attempt to compress extensive information on customer behaviour. Thus, a (visual) analysis – based on the *KDE* – makes it possible to identify 'dead spots' within a shopping centre. Furthermore, the spatially based analysis of customer routes simplifies the calculation of the *PBR*, which represents fundamental information for optimising rent payments by taking customer behaviour into account.

The regression of consumer frequency and the visual analysis show a significant negative correlation between passers-by-frequency and distance to *MC*. This confirms the results of Carter and Vandell (2005) and Carter and Haloupek (2000), regarding the decreasing passers-by-frequency when there is an increased distance to the *MC*.

### 4.4.3 Coupling

A GIS expands the possibilities for analysing coupling in terms of spatial aspects. First of all, a *GIS* allows us to represent coupling data graphically and hence supports visual and spatial analysis. The coupling values can be combined with several other forms of data and presented in a complex and condensed manner. The spatial database of our *GIS* facilitates the selection and visualisation of specific extracts from the total customer flow and consequently offers a wide range of possibilities for exploring the spatial behaviour of shoppers that underlies the actual coupling. Additionally, the coupling data can be combined with data about spatial relations between the shop units. This combination of a coupling matrix and a distance matrix offers a wide range of multivariate analysis methods. The analysis of customer flows and coupling yielded recommendations that might improve overall customer density and therefore raise revenue.

Focusing on shops of the same retail category, we can confirm the results of Brown (1992) that – regarding all retailers – a higher coupling for shops of the same retail category can be observed. At the same time, the coupling of the same retail category differs from retail category to retail category. While the category of fashion clearly shows a higher internal coupling, lower coupling for the same

retail category is observed for the categories of food and other goods. One reason is that this retail category – in contrast to fashion – generally yields a low coupling (Heinritz et al., 2003). Another reason is that within the sample, many more shops in the fashion category can be observed than those in categories such as food.

The hypothesis that the amount of coupling depends on the distance between two shop units could not be proved in our case. The data rather suggest a weak tendency for higher coupling between spatially separated retailers than between the closer ones. From the perspective of shopping centre management, this is a desirable condition, because a dispersed distribution of high-coupling retailers extends the walking paths of customers between these stores and thus offers more opportunities for coupling with other stores. Given that coupling is, inter alia, a function of the shops' retail categories, coupling and category clustering can be seen as two sides of the same coin. Thus, it should be possible to apply the results of Yuo (2010) – who detected a stronger clustering in more complex and multi-storey *SC* – to the case of coupling. The situation in the two-storey shopping centre of our investigation supports this perception, in view of the fact that a comparatively low level of clustering (compared to the Asian multi- storey centres analysed by Yuo) is associated with strong coupling, also over greater distances.

This paper provides the first attempt to analyse shopping centre passers-by-frequencies by modelling a multi-floor *GIS* network. The results can be assessed from two perspectives, namely methodological (use of *GIS*) and content (retail concentration, pass ratio, coupling).

Generally, the application of a GIS makes it possible to connect the non-spatial attribute data of a customer survey to a spatial geometry, using a polyline network. This geometry is the basis for almost unlimited ways of calculating spatial data (e.g. distances) within a shopping centre, which can be connected to customer shopping patterns – e.g. coupling or high-frequency locations. Hence, it is possible to identify 'dead spots' by using kernel density estimation or analysing coupling in dependence on shop-to-shop distance. At the same time, not only a generalised shopping centre concentration index, like the *Herfindahl index*, but also distance-based methods, such as the clumping method, *p-Median* or *Ripley's K*, can be used. Furthermore, *GIS* offers an automatic tool for detecting spatial autocorrelation by analysing the residual value – compressed to *Moran's I* – as well as tools for geographically weighted regression. Overall, the *GIS*-use within shopping centre research simplifies sophisticated spatial analysis and provides a basic tool for automating shopping centre research and thus creates an optimal tenant arrangement.

Although the paper demonstrates that the use of *GIS* may improve shopping centre research from both a technical and a content perspective, much remains to be done on implementing *GIS* into shopping centre research and into shopping centre management. In particular, the data survey must be organised in a more automated way, by using automatic traffic counters or customer tracking with *RFID*s. An issue which appears in the context of these applications – especially in Germany – is that of privacy policy. The analysis of results from the customer data should be considered in connection with the small sample of one shopping centre and whether sufficient standards have so far been achieved in spatial analysis. Thus, detecting a certain retail concentration depends not only on the spatial distribution of the shops, but also on the chosen parameters within the clumping method (clumping size and radius). The constitution of retail categories is another very important problem. Due to its statistical characteristics, the chosen *NACE* classification system – must be

regarded as an unrealistic reflection of the customer perspective. In this context, it would be very useful to set up a research retail classification system, which could then provide a basis for the comparing results from different articles.

Finally, a *GIS*-system can be regarded as a very effective way of analysing and optimising the tenant mix within (scientific) shopping centre research. This also applies within the area of shopping centre management, because it is the best and so far the only possibility for integrating spatial aspects – in an almost automated manner – into the analysis of customer shopping patterns. Nevertheless, it is crucial to verify the results of this paper by analysing a larger number of *SC* and by focusing not only on passers-by-frequency or coupling, but also on shopping centre rents.

## 4.5 References


Brown, S. (1992). Tenant Mix, Tenant Placement, and Shopper Behavior in a Planned Shopping Centre, *The Services Industries Journal*, 12(3), 384-403.

Carter, C. C., Allen, M. T. (2012). A method for determining optimal tenant mix (including location) in shopping centers, *Cornell Real Estate Review*, 10, 72-85.

Carter, C. C., Haloupek, W. J. (2000). Spatial Autocorrelation in a Retail Context, *International Real Estate Review*, 3(1), 34-48.

Carter, C. C., Haloupek, W. J. (2002). Dispersion of stores of the Same Type in Shopping Malls: Theory and Preliminary Evidence, *Journal of Property Research*, 19(4), 291-311.

Carter, C. C., Vandell, K. D. (2005). Store Location in Shopping Centers: Theory and Estimates, *Journal of Real Estate Research*, 27(3), 237-65.

Castle, G. H. (Ed.) (1998). *GIS in Real Estate – Integrating, Analyzing and Presenting Locational Information*, Appraisal Institute, Illinois.

Culley, J. (2010). *State of play of GIS usage in the Real Estate Industry*. Paper presented at the 17th Annual European Real Estate Society Conference, 2010, Milan.

Dawson, J. A. (1983). *Shopping Center Development*, Longman, London.

Des Rosiers, F., Thériault, M., Lavoie, C. (2009). Retail Concentration and Shopping Center Rents: A Comparison of Two Cities, *Journal of Real Estate Research*, 31(2), 165-207.

Dijkstra, E. W. (1959). A Note on Two Problems in Connexion with Graphs, *Numerische Mathematik*, 1, 269-271.

Dubin, R. A., Pace, R. K., Thibodeau, T. G. (1999). Spatial Autoregression Techniques for Real Estate Data, *Journal of Real Estate Literature*, 7(1), 79-95.

Ehlers, M., Schiewe, J. (2012). *Geoinformatik*, Wissenschaftliche Buchgesellschaft, Darmstadt.

Eurostat (Ed.) (2008). *Statistical Classification of Economic Activities in the European Community, Rev. 2*.

Heinritz, G., Klein, K., Popp, M. (2003). *Geographische Handelsforschung*, Gebrueder Borntraeger Verlagsbuchhandlung, Berlin.

Horni, A., Ciari, F., Axhausen, K.W. (2012). Coupling Customers' Destination Choice and Retailers' Location Choice. *MATSim*, *Arbeitsberichte Verkehrs- und Raumplanung*, 808, IVT, ETH Zürich.

Hoyler, M. (2005). Transnationale Organisationsstrukturen, vernetzte Städte: ein Ansatz zur Analyse der globalen Verflechtung von Metropolregionen, *Informationen zur Raumentwicklung*, 7, 431-438.

Klein, K. (2007). *GIS im Einzelhandel*, in: Klein, R., Rauh, J. (Ed.). Analysemethodik und Modellierung in der geographischen Handelsforschung, LIS-Verlag, Passau.

Lange, N. de, Kanzler, K., Plass, C. (2006). Erarbeitung eines Konzeptes zum Einsatz von Geoinformationstechnologien für die Stadt Lingen, Project study. Lingen.

Levy, M., Weitz, B., Grewal, D. (2013). *Retailing Management*, Mcgraw-Hill, New York.

Longley, P. A., Goodchild, M. F., Maguire, D. J., Rhind, D. W. (Eds., 2005). *Geographic Information Systems and Science*, Wiley & Sons, Hoboken.

Nakaya, T., Yano, K. (2010). Visualising crime clusters in a space-time cube: an exploratory data-analysis approach using space-time kernel density estimation and scan statistics, *Transactions in GIS,* 14(3), 219-377.

Nelson, R. L. (1958). *The Selection of Retail Locations*, F. W. Dodge, New York.

Okabe, A., Funamoto, S. (2000). An exploratory method for detecting multi-level clumps in the distribution of points: a computational tool, VCM (variable clumping method), *Journal of Geographical Systems*, 2(2), 111-120.

Okabe, A., Satoh, T., Sugihira, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool, *International Journal of Geographical Information Science*, 23(1), 7-32.

Prim, R. C. (1957). Shortest connection networks and some generalisations, *Bell System Technical Journal*, 36(6), 1389-1401.

R Core Team (2013). *R: A language and environment for statistical computing,* R Foundation for Statistical Computing, Vienna.

Roach, S. A. (1968). *The theory of random clumping*, Methuen, London.

Segerer, M. (2011). *Geographische Informationssysteme in der Immobilienwirtschaft. Praxiseinsatz und Konzeptionsmöglichkeiten*. Self-published by the IRE|BS Department of Real Estate, Regensburg.

Siebers, P.-O., Aickelin, U. (2007). A multi-agent simulation of retail management practices, *Proceedings of the 2007 Summer Computer Simulation Conference*.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Chapman & Hall, London.

Shiode, S., Shiode, N. (2009). Detection of multi-scale clusters in network space, *International Journal of Geographical Information Science*, 23(1), 75-92.

Stepper, M. (2014). *Stärkung der innerstädtischen Einzelhandelslagen vor dem Hintergrund des zunehmenden Online-Einkaufs,* in: Küpper, P., Levin-Keitel, M., Maus, F., Mueller, P., Reimann, S., Sondermann, M., Stock, K., Wiegand, T. (Eds., 2014). Raumentwicklung 3.0: Gemeinsam die Zukunft der räumlichen Planung gestalten. Akademie für Raumforschung und Landesplanung, Hannover.

Taylor, P. J., Walker, D. R. F. (2004). Urban Hinterworlds Revisited, *Geography*, 89(2), 145-151.

Thrall, G. I. (1998). GIS applications in real estate and related industries, *Journal of Housing Research*, 9(1), 33-59.

Walid, A., Moulin, B. (2006). How artificial intelligent agents do shopping in a virtual mall: A 'believable' and 'usable' multiagent-based simulation of customers' shopping behavior in a mall, in: Lamontagne, L., Marchand, M. (Eds.). Advances in Artificial Intelligence, 73-85, Springer, Berlin, Heidelberg.

Wells, V. K. (2012). Foraging: An ecology model of consumer behaviour?, *Marketing Theory*, 12(2), 117–136.

Xiea, Z., Yanb, J. (2008). Kernel Density Estimation of traffic accidents in a network space, *Computers, environment and urban systems*, 32(5), 396-406.

Yuo, T. S.-T. (2010). *Measurement of Retail Concentration and Variety in Vertically-used Large-scale Retail Properties*. Working Paper, ERES 2010, Milan. Reading: Working Papers in Real Estate & Planning 2/04, University of Reading Business School.

Yuo, T. S.-T., Crosby, N., Lizieri, C., McCann, P. (2004). *Tenant Mix Variety in Regional Shopping Centres: Some UK Empirical Analyses.* Reading: Working Papers in Real Estate & Planning 2/04, University of Reading.

Yuo, T. S.-T., Lizieri, C. (2013). Tenant Placement Strategies within Multi-Level Large-Scale Shopping Centers, *Journal of Real Estate Research*, 35(1), 25-52.

# 5 Variable Clumping Method and Mean-k-Nearest-Neighbor Method- Introducing two new approaches to retail concentration measurement to shopping center research

**Jens Hirsch** **Matthias Segerer**

**Abstract**

Planners and operators of shopping centers have to deal with two main challenges, namely the tenant mix (shops of specific retail categories) and the arrangement of tenants (*category concentration / clustering or dispersion*). The question of whether shops in the same retail category should be spatially concentrated or dispersed within a shopping center, has not yet been answered satisfactorily and there is reason to believe that it cannot in fact be answered. A first step in empirical research on this issue is the quantitative assessment of category concentration. Common methods of measuring the degree of clustering result in global measures, without identifying individual clusters at all or they lack theoretical foundation. This paper suggests the use of the *variable clumping method*, in order to consider the empirical situation of clustering within a shopping center. This enables the identification of statistically significant clusters on different spatial scales, as well as the pinpointing of shops actually constituting these clusters. The procedure and its results may be of importance for both practitioners and theorists. Academic researchers have tried to identify factors which influence the success of a shopping center. The degree of concentration or dispersion is one of these factors and can be analyzed in a more profound manner by means of the *variable clumping method*. The application of this method may enable further insights for research activities dealing with the coherence of retail category concentration and structural features of shopping centers. Planners and operators can use the methodology to analyze the situation in their centers and to identify potential improvements, according to scientific theories.

The paper provides a compendium of current research on retail concentration in shopping centers, explains the general principle of the *variable clumping method* and presents some sample results for a medium-sized German shopping center. Finally, with the *mean-k-nearest-neighbor method*, we introduce another new method which facilitates the analysis of concentration tendencies at a more abstract level. Both new methods are also applicable to urban retail agglomerations. The paper aims at initiating further research, including the effect of clustering on rents and customer-flow distribution.

## 5.1 Introduction

A considerable amount of research has been conducted on identifying success factors of shopping centers. These studies have focused mainly on specific determinants of rents, space allocation, types of lease, agency theory and the arrangement of stores (Carter, 2009). The analysis of store location in shopping centers deals with the arrangement of non-anchor stores and attempts to find optimization criteria that maximize revenue and sales. Concerning the arrangement of shops from different retail categories, the debate revolves around the advantages and disadvantages of concentration versus dispersion. Des Rosiers et al. (2009) state that category concentration – measured by the *Herfindahl index* – significantly and negatively affects shopping center rents. But their research only enables inferences on the general tenant mix, because the *Herfindahl index* does not represent the spatial arrangement of shops from different retail categories. Our paper does not aim to find an ideal shop arrangement (clustering vs. dispersion), but to develop a reliable method for measuring the spatial concentration of retail categories. We therefore apply the 'theory of random clumping' – outlined by Roach (1968) – on the tenant arrangement within a shopping center. The suggested approach is a further development of *the variable clumping method*, which we adapted to the special situation in a shopping center. It features a reliable technique for integrating the spatial concentration of retail categories in shopping center research and is able to consider different spatial scales. We further introduce the *mean-k-nearest-neighbor method* as an easy-to-implement method, which can detect clusters in a shopping center by analyzing mean distances of shops to a varying number of nearest neighbors.

With respect to this methodological focus, the paper is organized as follows. Section 5.2 provides an overview of the literature on retail concentration in shopping center research, Section 5.3 presents the theoretical framework and some empirical results of the *variable clumping method* and Section 5.4 depicts the *mean-k-nearest-neighbor method*. Section 5.5 concludes with an overview of the most important results and further research questions concerning retail concentration and the application of the two introduced methods.

## 5.2 Literature Review: Concentration of Retail Categories

According to Nelson (1958), tenant mix and arrangement are not only aligned with synergetic effects (shared and suspicious business), but also with competition between the tenants and are consequently intended to be optimized by shopping center operators. Depending on the tenants' retail categories, there is a more or less strong potential for reciprocal coupling. Anchor stores attract a large number of customers and affect the most important flows within the center. The aim of a shopping center operator is to maximize these flows and to create the highest level of coupling with other stores. Yuo et al. (2004) identify six factors which classify the shopping center tenants by using a factor analysis. Based on this classification, they find that a high share of 'core' retail categories – such as Fashion, Comparison Variety, Selective Information and Health – has a positive impact on shopping center rent. Customers aim to minimize their efforts to get the goods they want – as long as the shopping trip is a rational affair. Under these circumstances, customers prefer centers with a large supply of the abovenamed core retail categories. If customers seek to minimize their efforts to visit the shops of certain retail categories, they should also prefer a clustered arrangement of shops from the same retail category within a shopping center. Like Yuo et al. (2004) Des Rosiers et al. (2009) use the *Herfindahl index* to measure the concentration of retail categories, but demonstrate a

negative correlation between retail category concentration and rent. In the studies of these authors, 'concentration' measured by the *Herfindahl index* depicts only the partition of sales area between several retail categories and does not detect actual spatial clusters. In contrast, Carter and Haloupek (2002) use a model of spatial correlations based on the p-median approach. Their findings were that non-anchor stores in the same retail category should be located in a dispersed pattern within a shopping center. Yuo (2013) argues that the p-median approach used by Carter and Haloupek (2002) *"[…] may not be a proper concept to explain the dispersion of retail stores of the same type […]"*, because of the comparative nature of retail facilities in one category. In his investigation of vertically-used retail properties, he discovered an interesting coherence between the degree of clustering and customer shopping efforts resulting from the structural features of a shopping center. He showed – using a modified *Herfindahl index* in combination with the *InterConnection* approach according to O'Neill (1991) – that the degree of clustering rises with an increasing number of floors and the complexity of the center's floor plan. The results of Yuo (2013) suggest that the question of how to arrange the shops in a retail category within a shopping center cannot be answered generally. In fact, the optimal strategy seems to depend not least on structural characteristics of the center. From a methodological perspective, Yuo (2013) does consider spatial aspects in his approach, but the applied method is unable to differentiate between 'low' and 'high' retail concentration, because he only considers the total number of shops that are part of single clusters, consisting of at least 4 stores from the same retail category within 5 meters. This means that two clusters consisting of four shops in the same retail category are regarded as equally significant in terms of retail concentration, as one cluster consisting of eight shops. However, the probability of observing an eight-shop cluster within a shopping center is much lower than of observing two four-shop clusters. In short, the existing tools for measuring the concentration of retail categories within a shopping center either do not consider spatial aspects at all – like the generalizing *Herfindahl index* – or if they do consider spatial aspects – like the *p-median model* or the *InterConnection* approach – they do not depict the customers' maximizing criterion realistically and do not consider the stochastic nature of the measures.

## 5.3 Variable Clumping method

### 5.3.1 Theoretical Framework

Given the deficits of existing approaches, it is necessary to find a method which enables modeling both the spatial distribution of shop clusters, as well as cluster size, measured with a probability-based concept and to apply it to the measurement of retail concentration within a shopping center. One approach that accounts for both criteria is the *variable clumping method* developed by Okabe and Funamoto (2000)

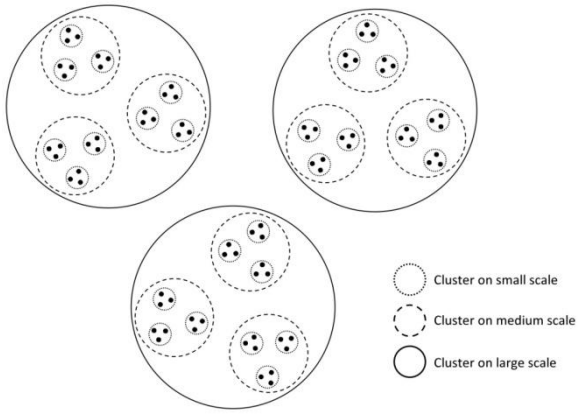
### 5.3.2 The idea of the variable clumping method

This represents an alternative approach, which considers both criteria by identifying clusters on a variable scale. The idea of the *clumping method* was first outlined by Roach (1968). Okabe and Funamoto (2000, p. 112) describe the method and define a clump as follows:

> *"It is a class of methods for finding 'clumps' in the distribution of points,* $p_1; \ldots; p_n$ *over a bounded region* $S$*. Usually a clump is defined in terms of circles centered at given points* $p_1; \ldots; p_n$ *[…].The radius of the circles is called a clump radius. A clump is then defined as a set*

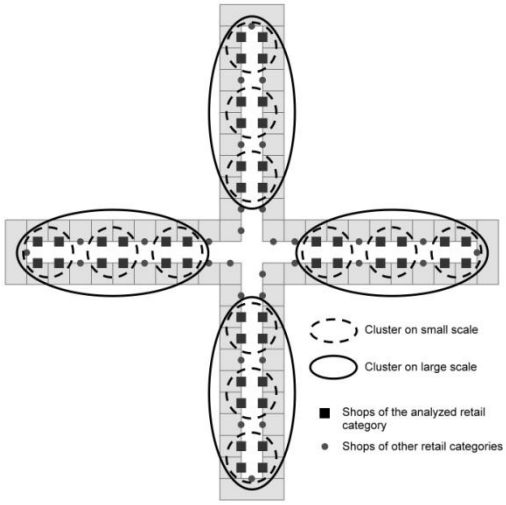
*of points whose circles are connected. The number of connected circles in a clump is called the size of the clump"*

**Figure 1: Idealized situation of multi-level clumps**




*Source: Adapted from Okabe and Funamoto, 2000, p. 112.*

**Figure 2: Multi-level clumps within a shopping center**




*Source: Own Design 2011.*

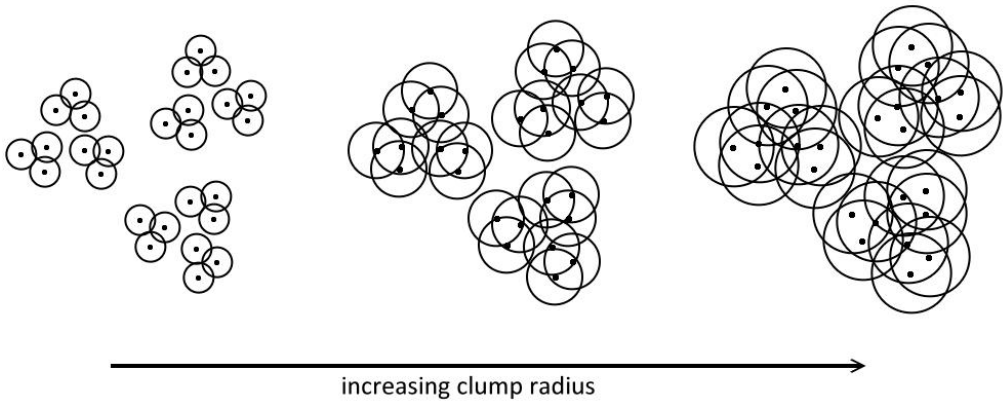
The idea of the clumping method is to draw circles with the same (clumping) radius $r$ around each point of a given point distribution and to analyze the emerging pattern of overlap. In order to deal with networks or the situation in a shopping center, these circles are denominated more generally as *buffers.* The buffers of two points overlap if their distance from each other is smaller than $2 \cdot r$. Depending on the distribution of points, a more or less high number of buffers will overlap. The number of points and buffers respectively comprising a clump is called the *clumping size* $i$. The result of the clumping process is called the *clumping state* $C$, which is characterized by the number of points n, the used clumping radius r and the number $N(i|r)$of observed clumps with a certain size $i$ (see Okabe and Funamoto 2000; Shiode and Shiode 2009). Consequently, the clumping state can be written as follows:

$$C(r) = N(1|r), N(2|r), \dots N(n|r) \qquad (1)$$

The *clumping-state C* summarizes the number of clumps of different sizes. The *clumping-state* is varied by modifying the *clump radius r,* and the *variable clumping method* enables analyzing clumps on a variable scale based upon a continuous variation of *r* (see Figure 1). If $r$ reaches a critical value $r_{max}$, which depends on the actual distribution, all points will form one single cluster of size $n$. If the clump radius is varied from zero to $r_{max}$, the number of larger clumps increases. The comparison of the observed *clumping state* $C^O$ with the state expected under a random distribution $C^R$ is the precondition for identifying significant clumps. "*One way to compare the clump states C of both distributions for a range of clump radii is to change r in the increment of a unit length."* (Shiode and Shiode 2009, p. 80). This *unit length* is called $L$. The clump radii of the individual steps of the analysis are $r_h = L \cdot h; h = 1, 2, \dots q$. The radius $r$ increases up to $r_h = L \cdot q$ and the resulting clumping states are compared with the empirical one. A clump can be classified as significant, if the number of clumps of size $i$ for radius $r$ is significantly larger than the number of clumps that would appear in the distribution of random points.

Applied to 'retail concentration', a shop is regarded as 'clustered' if it is part of a significant clump. According to Shiode and Shiode (2009), the variable clumping method can be accomplished by the course of action shown in Figure 3, which has been adapted to the case of stores within a shopping center.

**Figure 3: Clumping states for different clump radii**




*Source: Okabe and Funamoto, 1999, p. 114.*

The selection of clumping radius involves a change in the clumping state (see Figure 3). The clumping states for the three clump radii in Figure 4 are:

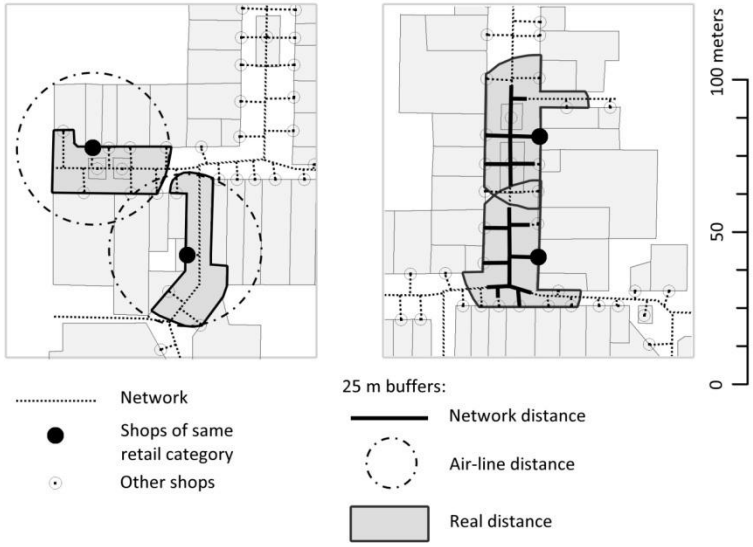
$$C(r_1) = \{N(1|r_1) = 0, N(2|r_1) = 0, N(3|r_1) = 9, N(4|r_1) = N(5|r_1) = \ldots = N(27|r_1) = 0\}$$
$$C(r_2) = \{N(1|r_2) = N(2|r_2) = \ldots = 0, N(8|r_2) = 0, N(9|r_2) = 3, N(10|r_2) = \ldots = N(27|r_2) = 0\} \quad (2)$$
$$C(r_3) = \{N(1|r_3) = N(2|r_3) = \ldots = N(26|r_3) = 0, N(27|r_3) = 1\}$$

### 5.3.3 The variable clumping method and retail concentration within a shopping center

Generally, the *VCM* analyzes the distances between shops in a certain retail category and searches for significant clumps. The number and size of these clumps represent the observed concentration within a certain retail category. Given that the floor-space of many shopping centers (including the one in our study) has notches (that means, that the polygon representing the center is concave), the representation of walking distances between shops cannot not be achieved by simply drawing circles around the shops (see below). This would correspond to the approach of Okabe and Funamoto (2000), which model the radius $r$ using air-line distances. Shiode and Shiode (2009) apply this approach to networks. However this approach would basically be better suited to consumer behavior – especially in complex scaled shopping centers – but in our opinion, this method is still not optimal. In order to differentiate between the two approaches, Shiode and Shiode (2009) refer to them as the *Planar Variable-Distance Clumping Method (PL-VCM)* and the *Network-Based Variable-Distance Clumping Method (NT-VCM)*. In *PL-VCM*, points can be distributed universally inside a given region and buffers can be drawn using air-line distances. *NT-VCM* restricts the location of analyzed points to a given network and uses network distances to create buffers. The case of a shopping center can be considered as a hybrid. Applied to retail concentration within a shopping center, each shop entrance can be regarded as a point $p_i$ which is located on the edge $E$ of a bounded region $S$ represented by the shopping center floor space (see Figure 2). The movement of customers is not restricted to any given network. Instead the customers are assumed to be able to move freely within the floor-space

from one shop entrance to another, along the shortest path between them. Figure 4 demonstrates that both air-line distance and network distances would yield a distorted picture of the situation in our specific case with freely moving customers inside a region with notches. The application of air-line distances can result in an overestimation of clumping, whereas the use of network distances might not detect clumps where they actually exist, if a simple network consisting of one main line and a branch to every shop is used. A network that connects all shops with each other would be necessary, in order to avoid this problem. As *VCM* uses *Monte Carlo simulation* to distribute (see below) shops randomly: there cannot be only one network, but a separate one for each simulation run (done in our analysis). The situation would become less complicated if the shops' physical locations were considered as a given fact and only the allocation of tenants (representing specific retail categories) had to be determined. This latter approach corresponds more to the perspective of center management, whereas that of distributing the shops alongside the complete edge of the floor-space corresponds to the developer perspective. The explanations in this paper describe the latter perspective, which is associated with an increased calculation effort, but the principal methodology can easily be transferred to the case of fixed store locations.

**Figure 4: Comparison of clump appearance by using different measures of distance**




Okabe and Funamoto (2000) and Shiode and Shiode (2009) suggest computing the *Monte Carlo*-generated random clump states by generating *Voronoi diagrams*, performing a *Delaunay triangulation* and deriving the *minimum spanning tree.* The *minimum spanning tree* is the shortest graph that connects all points, and enables an easy and fast computation of the number of clumps. In our case of an irregularly shaped and concave boundary, we prefer the computation of the *minimum spanning tree* by applying the *Prim algorithm* (Prim, 1957). Okabe and Funamoto (2000) also mention this possibility and state that the computational time is the same as that of using *Voronoi diagrams*.

**Figure 5: Process of detecting multi-scale clumps within a shopping center**

**0. Initial settings: Decide on the dimension of *unit length (L)*, *steps (q)* and $m_{sim}$**

Despite the high amount of necessary computing power and the introductory nature of this analysis, the parameters were chosen comparatively high ($m_{sim}$ and $q$) respectively low ($L$): $L$ = 5m, $q$ = 10, $m_{sim}$ = 10,000 (we chose $L$ = 1m, $q$ = 50, $m_{sim}$ = 1,000 for a second analysis)

**1.Clumping of the observed distribution with clumping state $C^O(r)$ for all radii to explore**

Calculate $C^O(r_h) = \{N^O(1|r_h), \dots N^O(n|r_h)\};\ h = 1, 2, \dots q;\ r_h = L * h;\ n = 'number\ of\ elements'$

**2. Monte Carlo simulation**

Generate $n$ randomly distributed points on the edge $E$ of the shopping center mall. From a practical point of view, this approach assumes that size and shape of the mall are already known, but the layout of shop units (and the allocation of tenants from different retail categories) is still amenable (developer perspective). An analogous approach, where store locations are defined, but tenants / retail categories still have to be assigned (management perspective) is feasible by generating *n* points randomly on a given set of locations.

**3. Clumping of the generated random distribution with clumping state $C^R(r)$**

Calculate $C^R(r_h) = \{N^R(1|r_h), \dots N^R(n|r_h)\};\ h = 1, 2, \dots q;\ r_h = L * h;\ n = 'number\ of\ elements'$

**4. Iteration of steps 2 and 3 $m_{sim}$ times**

The results is a set of clumping states $C_u^R(r_h);\ h = 1, 2, \dots q;\ r_h = L * h;\ u = 1, 2, \dots m_{sim}$

**5. Sorting the clump states of the random distribution**

Sort the clumping states $C_u^R$ for each $r_h$ in ascending order

**6. Identification of critical value**

This is the crucial step in determining the marginal value for significant clumps.
Determine:
$C_\alpha^R(r_h) = \{N_\alpha^R(1|r_h), \dots N_\alpha^R(n|r_h)\}$ where
$H_0: N^O(i = k|r_h = L * h) \leq N_\alpha^R(i = k|r_h = L * h)$ at an upper critical value of $N_\alpha^R(i|r_h)\ with\ \alpha \in \{0.01, 0.05, 0.1\}\ \forall\ (k \in \{1,2, \dots n\}\ \wedge\ h \in \{1,2, \dots q\});$

**7. Hypothesis testing: The observed clumping state has to be compared with the critical values gained from step 6.**

If $N^O(i|r_h) \geq N_\alpha^R(i|r_h); i = 1, 2, \dots n; h = 1, 2, \dots q;$ then $H_0$ can be rejected on the given level of significance and the points that are contained in $N^O(i|r_h)$ compose a significant clump.

*Source: Adopted from Shiode and Shiode, 2009.*

#### 5.3.3.1 Technical implementation

The main computation was done using the statistical programming language *R* (R Core Team, 2013), which enables the calculation of the necessary geometric items and processes, by applying simple scalar and vector analysis. The initial data collection and preparation was with *GIS ESRI ArcGIS,* which enabled an easy digitalization of the shopping center's georeferenced floor plan. We used *ArcGIS* to create three separate files that were applied as input for the R software, in which the main calculations for the variable clumping method were conducted. These three files contain the coordinates of

(1) the shopping center area where the customers can move between the shops (*mall*)
(2) the shops, including their respective retail category
(3) all concave vertices of the mall-polygon.

These concave vertices are important, because they can be part of the shortest path between two shops, which has to be calculated as part of Prim's algorithm. It is obvious that convex vertices are not candidates for any shortest path.

In the following analysis, we describe the calculation of the clumping state of the observed / empirical distribution of shops from a certain retail category. Subsequently, we outline the attribution of significant clumps.

First of all, we create all possible vectors connecting the shops with each other, the concave vertices with each other, and the shops with the concave vertices. Those connections now have to be deleted, which are not completely within the mall. Therefore a check is run to determine whether they intersect with any of those lines enclosing the mall, or whether their midpoint lies outside the polygon. The latter is necessary, in order to remove those connections lying completely outside the mall. This is the only operation which was conducted with the help of an existing function of *R*, called *point.in.polygon()* (the function *point.in.polygon()* is based on the *C* function *InPoly()* developed by O'Rourke, 1998). The remaining connections represent all possible segments of the shortest paths between the given shops.

Subsequently, the minimum spanning tree is obtained by applying the *algorithm of Kruskal*. This also requires the implementation of *Dijkstra's algorithm*, in order to find the shortest path between two shops, based on the above created set of possible connections. Finally, the resulting minimum spanning tree can be evaluated with regard to the existence of clumps for any given clump radius $r_h = L * h$ as decribed by Shiode and Shiode (2009). The idea behind this evaluation is that two shops form a clump, if the distance between them is does not exceed twice the clump radius. The number of clumps of certain sizes provides the empirical clumping state.

The theoretical clumping state, which can be expected under a random distribution of shops, is obtained via *Monte Carlo simulation* ($m_{sim}$*=10,000*). Therefore, we distribute *N* shops (the number corresponds to the number of shops in the specific retail category) randomly alongside the edge of the mall and compute the clumping state using the procedure described above. This is done *n* times, by which the distribution of clumping states is obtained. As a result, we might, for instance, state that there are on an average 2.6 clumps of size 3 (shops) for a clump radius of 20 meters, including the

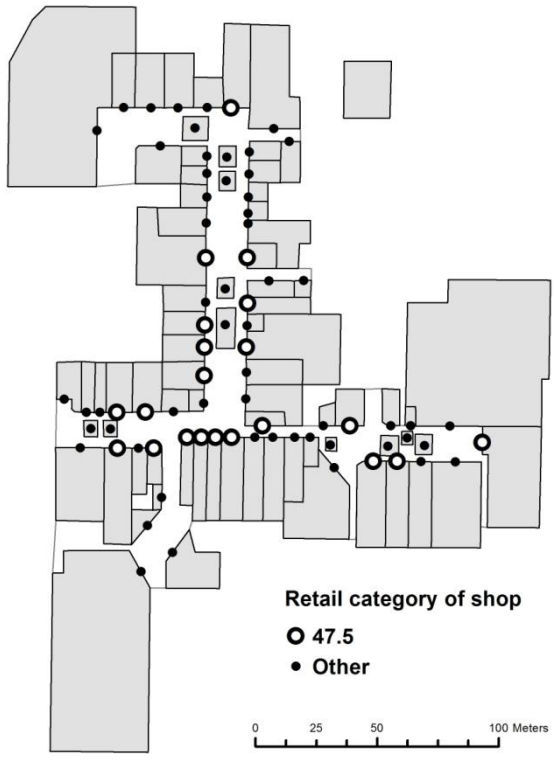
variance of this value. If we assume a significance level of 5%, the number of clumps of size 3 for a radius of 20m in the empirical distribution must exceed the 95% quantile of the respective clumps from the *Monte Carlo simulation*.

### 5.3.4 Results of the variable clumping method for an sample shopping center

#### 5.3.4.1 Data

The variable clumping method described in Section 5.3.1 was applied to the *NACE* retail category 47.5 ('Retail sale of other household equipment in specialized stores including textiles, electrical household appliances, furniture, lighting equipment and others') in one floor of a medium-sized German shopping center. Altogether, there are 66 shops in the analyzed floor, 21 of them category 47.5 shops (see Figure 6). The floor has a *Y*-shape with large anchor stores at the end of each arm. Obviously, there is a high concentration of 47.5 shops in the southern part of the center, with a conspicuous spot with four shops directly next to each other. The estimation of critical values for $N_{\alpha}^{R}(i|r_h)$ shows whether these apparent clusters are really statistically significant or purely a result of a random distribution. Table 1 shows the number $N^{O}(i|r_h)$ of observed clumps with size *i* with a given radius $r_h$. In order to test the significance of these empirical clumps, 21 elements were located randomly along the edge of the mall in every step of the *Monte Carlo*-analysis. The clumping states were computed for these $m_{sim}$*=10,000* random distributions and critical values were calculated.

**Figure 6: Stores in the analyzed shopping center**




#### 5.3.4.2 Estimation results

In contrast to the first visual impression, the comparison of the observed values (Table 1) with the critical values derived from *Monte Carlo simulation* at a 99%-level of significance shows that there is a significant clump of size 9 for a radius of 10m, formed by those shops in the south and south-west of the center (see Figure 7). Further clumps that are significant on this level are one of size 16 with a

clump radius of 15m, and one of size 20 with a clump radius of 20m. The high level of significance for these large clumps indicates a concentration tendency on a relatively large scale, and indeed, most of the shops are located in the southern half of the center (only one in the northern part). Those four shops located directly next to each other in the south of the center only form a clump of size four at a significance level of 90% (clump radius of 5m). It is surprising, that the clump of size 7 with a radius of 10m has only a significance level of 90%, but the *Monte Carlo simulation* shows that 5.9% of all simulations are associated with a clump of size 7 for this radius.

Table 2 shows the results of the variable clumping method for all retail categories. Some do not show any significant clumping at all or only at specific clump radii. The strongest clustering tendency is found for the categories 47.2 (food, beverages and tobacco in specialized stores), 47.5 (other household equipment in specialized stores*)* and 47.6 (cultural and recreation goods in specialized stores*).*

To make these results usable for an application in regression models that estimate the effect of certain factors on turnover or rent, some derived measures can be useful. These might, for example, be the share of clustered shops (concerning specific retail categories or all shops) or the corresponding sales area. To be able to differentiate between, for example, one cluster of size eight and two clusters of size four, the sales areas of each cluster should be weighted by the ratio of clumping size and total number of shops.

**Table 1: Observed clump states and levels of significance for retail category 47.5**

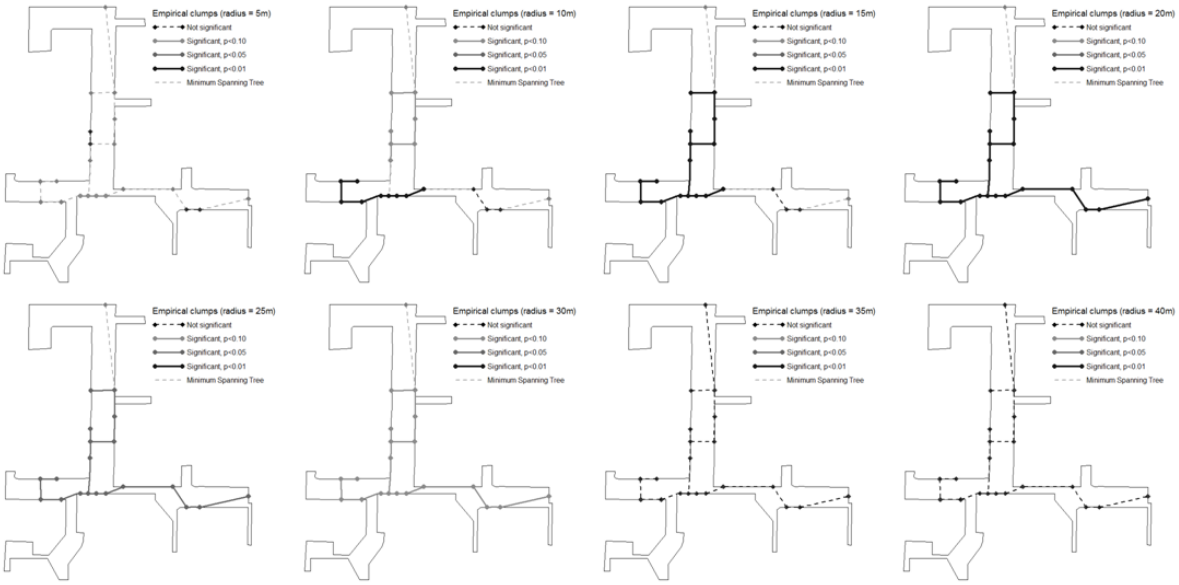
| **Number of Clumps** | *Clump size i* | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Clump radius $r_n$ [m]* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* | *17* | *18* | *19* | *20* | *21* |
| *5* | 13 | 2 | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *10* | 2 | 0 | 1 | 0 | 0 | 0 | **1** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *15* | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 | 0 | 0 | 0 | 0 |
| *20* | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| *25* | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| *30* | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | 0 |
| *35* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *40* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *45* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *50* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

not significant | **significant at p<0.10** | **significant at p<0.05** | **significant at p<0.01**

**Table 2: Results of the variable clumping method for all retail categories**

| ***Retail category*** | ***Number of shops*** | ***Significant clumps*** *("number of clumps" x "clump size" @ "clump radius"; ***: p<0.01, **: p<0.05, *:p<0.1)* |
|---|---|---|
| *47.1* | *3* | no significant clumps |
| *47.2* | *13* | **1x6**@10m**, **1x6**@15m*, **1x8**@35m* |
| *47.4* | *2* | no significant clumps |
| *47.5* | *21* | **1x4**@5m* , **1x7**@10m*, **1x9**@10m*** , **1x16**@15m*** , **1x20**@20m*** , **1x20**@25m** , **1x20**@30m* |
| *47.6* | *6* | **2x2**@10m*, **2x3**@25m**, **2x3**@30m**, **2x3**@35m**, **2x3**@40m*, **2x3**@45m**, **2x3**@50m** |
| *47.7* | *11* | **1x11**@30m** |
| *53.1* | *1* | --- |
| *56.1* | *2* | no significant clumps |
| *64.1* | *1* | --- |
| *79.1* | *3* | **1x2**@5m** |
| *95.2* | *1* | --- |
| *96.0* | *2* | no significant clumps |

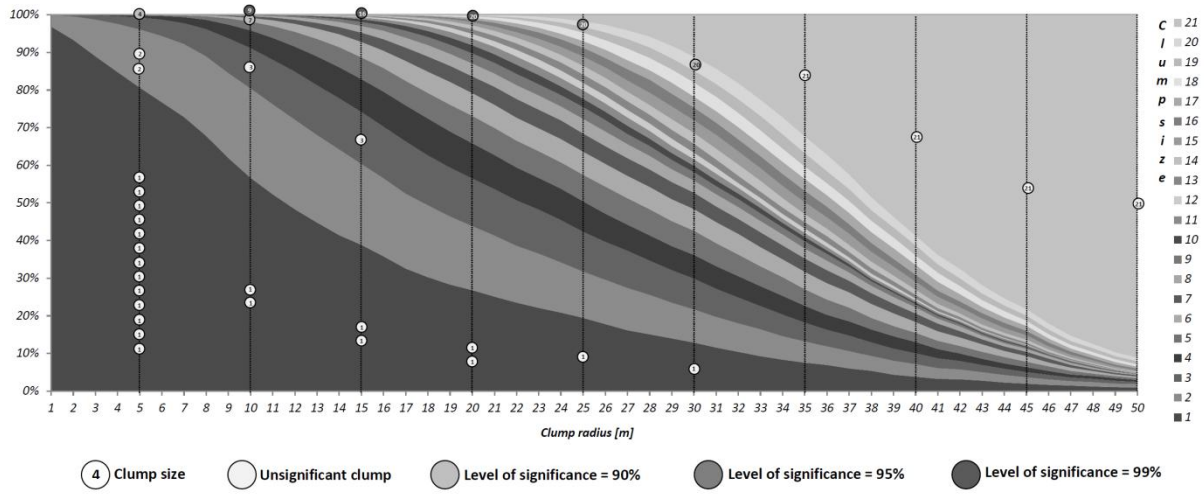
**Figure 7: Number of clumps for retail category 47.5 and clump radii from 5m to 40m**




The effect of changing *clump radii* can best be shown if the radius is altered in smaller steps then in the statements so far. Therefore, we gradually increased the *clump radius* from 1m to 50m in steps of 1m, conducted a *Monte Carlo simulation*, as described above and plotted the averagely occurring share of clumps with a certain size (see Figure 8, including the position of those clumps in the empirical distribution of shops in the investigated shopping center). As one might expect, the number of small clumps (*clumping size* = 1 represents isolated shops which are not part of any clump) decreases with a growing *clump radius*, while more large clumps occur. This figure can be also interpreted concerning the significance of the empirical clumps: This is high if the probability of occurring in the random distribution is low (corresponding to a narrow area in the figure) and vice versa. The previous comments have stressed the aspect of concentration and looked for empirical distributions with significantly many clumps. But Figure 8 can also be read the other way round. With smaller clumps of isolated shops becoming less probable for higher *clump radii*, the occurrence of these small clumps can also be interpreted as an indication of separation or dispersion. For example, if we were to find a clump of size 1 for a clump radius of 50m (which is not the case in the analyzed center), this single shop can be regarded as significantly isolated.

The *variable clumping method* can be further modified by considering the sales area of shops to give those shops with high sales area more weight. A conceivable method to achieve this weighting, would be to modify the *clump radius* for each shops dependent on its sales area, for example by a factor that represents the ratio of its sales area to the average value of all shops being analyzed. The *Monte Carlo simulation* has to distribute shops with the factual sales areas randomly which implies a higher number of simulations that are necessary to obtain stable results.

**Figure 8: Share of clumps of a certain size for continuous clump radii from 1m to 50m in steps of 1m, according to Monte Carlo simulation (circles: empirical clumps with number of shops and level of significance)**




## 5.4 Mean-k-Nearest-Neighbor

Not least because the variable clumping method needs considerably high computational power, we propose another method that takes into account retail category concentrations within a shopping center. We refer to it as the *mean-k-nearest-neighbor method* and give some examples of its application, with the example of a plain rectangular shopping center.

The idea of using distance measures in shopping center research is not new. Carter and Haloupek (2000) introduced several distance measures, including distances to the next store of the same type, and analyzed their impact on rents (see Hirsch et al., 2016, p. 4f, for further distance-based approaches which were used in the literature). Hirsch et al. (2016, pp. 15ff) analyze the mean distance between shops of the same retail category and between shops of other retail categories. If the mean distance of all shops in a certain retail category to all other shops is higher than the intra-category mean distance, the shops of this category seem to be spatially concentrated. This method delivers some initial indications of clustering, but it is neither able to pinpoint single clusters, nor does it consider the stochastic character of the analyzed issue.

Let's consider a circle with radius *r* as an easy example. The shops are situated along the edge of the circle, and the mean distance between two shops which are randomly distributed on the edge is $\bar{d} = {}^{4\pi}/_{r} = 1.274\, r$ (Mathai, 1999, p. 178). If we consider *n* points, which are randomly distributed on the circle, the mean distance of all shops remains this value, because each newly added point still has the above-mentioned mean distance to every point which already exists. To make the illustration more applicable to the situation of a shopping center, let us consider a rectangular shape of length 100m and height 20m instead of a circle. A *Monte Carlo simulation* ($n_{sim}$=10.000.000) yields a mean distance of 41.8m. This value should be taken into consideration if, for example, intra-category mean distances or other measures are analyzed.

The idea behind the *mean-k-nearest-neighbor method* is that the distribution of shops affects the mean distance to each shop's *k* nearest neighbors. If, for example, fives shops are located very close to each other, whereas the others are located in a dispersed pattern, the mean distance to the 5

nearest neighbors will be significantly higher than to the 4 nearest neighbors (see Figure 9d). We computed these *mean-k-nearest-neighbor distances* in the rectangular shopping center described above, for five different generic distributions of 14 shops. Thus, we obtained the mean distances for 1, 2… 13 nearest neighbors for each of the five distributions and denote this as the *mean-k-nearest-neighbor spectrum*. In addition, this spectrum was calculated for 10.000 random distributions in order to obtain the respective average values. Figures 9a-d show the resulting spectra of the four generic distributions (top left), the alteration of the different values of *k* (top right) and a map of the particular distribution of shops (bottom).

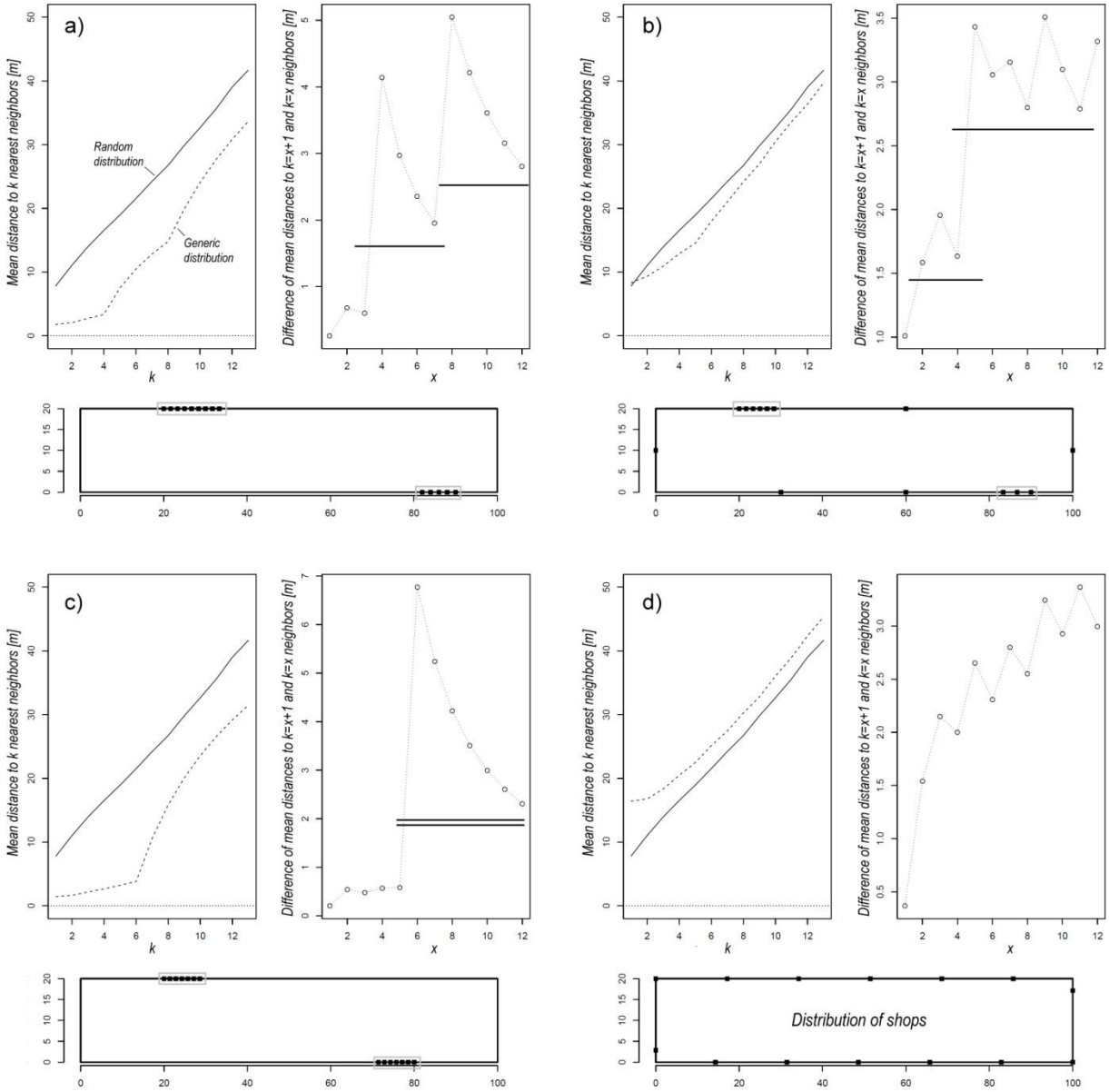
The situation where all shops are part of two separated clusters, one of 9 shops and one of 5 shops, is depicted in Figure 9a. The *mean-k-nearest-neighbor spectrum* features two obvious knees at *k=4* and *k=8*, where the mean distance increased distinctly more than for the random distribution. As described above, the *mean-nearest-neighbor distance* increases particularly strongly at these positions, because the 4 (8) nearest-neighbor-distance of the 5 (9)-shop-cluster are measured within the cluster, whereas the $5^{th}$ nearest neighbor lies further away. If we consider the difference between those mean distances for *k* and *k+1* nearest neighbors (Figure 9a, top right), it is even easier to identify the critical values of *k* indicating clusters (marked with horizontal lines). Figure 9b shows a similar case with two clusters (3 shops and 6 shops) and some dispersed shops. It can be interpreted in an analogous manner, as Figure 9a. In Figure 9c, the particularly strong increase in the mean distance from *k=6* to *k=7* indicates the existence of two clusters of seven shops. Finally, Figure 9d shows the results of *mean-k-nearest-neighbor method* for a uniform distribution of shops along the edge of the center. The mean distances are higher than for the random distribution for every value of *k* and show no obvious knees.

The results of the *mean-k-nearest-neighbor method* allow a quick and easy computation and initial assessment of clustering tendencies within a shopping center. In the case of floor plans which are more complex than the rectangular center analyzed before, the *mean-k-nearest-neighbor method* requires the application of techniques to determine the shortest paths (*Dijkstra's Algorithm*), just as in the case of the *variable clumping method*. The method should be complemented with other methods such as that of *variable clumping*, so as to obtain more significant results. Further research should analyze the ability of the method to identify clusters in more complex-shaped geometries or to develop a more sophisticated mathematical understanding of the results beyond a mere graphic evaluation.

## 5.5 Conclusion

This paper introduces a probability-based analysis of retail concentration into the shopping center research. In this context, the *variable clumping method* enables the detection of retail concentration within shopping centers at a multi-scale level using *GIS*-analysis. In contrast to previous investigations which mainly used the *Herfindahl index* as a global retail concentration measurement, the *variable clumping method* is able to identify significant clusters with varying spatial scales. The share of clustered shops or of clustered sales areas can be derived as more global measures and used for further analyzis. The paper summarizes the methodology for an implementation of the *variable clumping method* and gives some further indications of the characteristics and potential of the method.

**Figure 9a-d: Results of mean-k-nearest-neighbor method for different distributions of shops**




From a content perspective, the sample of the examined shopping center should be extended to obtain a reliable application for shopping center research. From a methodological perspective, the application shown above of the *variable clumping method* provides a sophisticated approach to identifying retail concentration within a shopping center. Consequently, it represents a basic tool for generating more reliable data of retail concentration, which are crucial for further analysis of optimal tenant placement and thus of rent optimization within shopping center research.

The paper further introduces the *mean-k-nearest-neighbor method* into shopping center research and provides some basic examples of its application. This method constitutes a straightforward approach to searching for clusters of certain retail categories. Further research should analyze the behavior of *mean-k-nearest-neighbor-spectra* for more complex floor plans and develop a better understanding of the method's statistical characteristics.

The application of the two newly introduced methods is not limited to the application in shopping centers. The analysis of urban retail agglomerations constitutes a similar field of interest and could benefit of the new methods in the same way.

## 5.6 References


Carter, C. C. (2009). What We Know About Shopping Centers, *Journal of Real Estate Literature*, 3(2), 165-180.

Carter, C. C., Haloupek, W. J. (2000). Spatial Autocorrelation in a Retail Context, *International Real Estate Review*, 3(1), 34-48.

Carter, C. C., Haloupek, W. J. (2002). Dispersion of stores of the Same Type in Shopping Malls: Theory and Preliminary Evidence, *Journal of Property Research*, 19(4), 291-311.

Des Rosiers, F., Thériault, M., Lavoie, C. (2009). Retail Concentration and Shopping Center Rents: A Comparison of Two Cities, *Journal of Real Estate Research*, 31(2), 165-207.

Dijkstra, E. W. (1959). A Note on Two Problems in Connexion with Graphs, *Numerische Mathematik*, 1, 269-271.

Hirsch, J., Segerer, M., Klein, K., Wiegelmann, T. (2016). The analysis of customer density, tenant placement and coupling inside a shopping centre with GIS, *Journal of property research*, forthcoming.

Nelson, R. L. (1958). *The Selection of Retail Locations*, F. W. Dodge, New York.

Mathai, A. M. (1999). *An introduction to geometrical probability. Distributional aspects with applications*, Gordon and Breach, Newark.

Okabe, A., Funamoto, S. (2000). An exploratory method for detecting multi-level clumps in the distribution of points: a computational tool, VCM (variable clumping method), *Journal of Geographical Systems*, 2(2), 111-120.

O'Neill, M. J. (1991). Evaluation of a conceptual model of architectural legibility. *Environment and Behavior*, 23(3), 259-284.

O'Rourke, J. (1998). *Computational Geometry in C*, 2nd Edition, Cambridge University Press.

Prim, R. C. (1957). Shortest connection networks and some generalisations. *Bell System Technical Journal*, 36(6), 1389-1401.

R Core Team (2013). *R: A language and environment for statistical computing,* R Foundation for Statistical Computing, Vienna.

Roach, S. A. (1968). *The theory of random clumping*, Taylor & Francis, London.

Shiode, S., Shiode, N. (2009). Detection of multi-scale clusters in network space, *International Journal of Geographical Information Science*, 23(1), 75-92.

Yuo, T. S.-T., Crosby, N., Lizieri, C., McCann, P. (2004). *Tenant Mix Variety in Regional Shopping Centres: Some UK Empirical Analyses.* Reading: Working Papers in Real Estate & Planning 2/04, University of Reading.

Yuo, T. S.-T., Lizieri, C. (2013). Tenant Placement Strategies within Multi-Level Large-Scale Shopping Centers, *Journal of Real Estate Research*, 35(1), 25-52.

# 6 Do urban tourism hotspots affect Berlin housing rents?

**Philipp Schäfer** **Jens Hirsch**

**Abstract**

**Purpose:** The aim of the study is to analyze whether urban tourism affects Berlin housing rents. Urban tourism is of considerable economic importance for many urban destinations and has developed very strongly over the last few years. The prevailing view is that urban tourism triggers side-effects, which affect the urban housing markets through a lack of supply and increasing rents. Berlin represents Germany's largest rental market and is particularly affected by growing urban tourism and increasing rents.

**Design / methodology / approach:** The paper considers whether urban tourism hotspots affect Berlin´s housing rents, using two hedonic regression approaches, namely conventional *ordinary least squares* (*OLS*) and *generalized additive models* (*GAM*). The regression models incorporate housing characteristics as well as several distance-based measures. The research considers *tourist attractions*, *restaurants*, *hotels* and *holiday flats* as constituents of tourism hotspots and is based on a spatial analysis *using geographic information systems* (*GIS*).

**Findings:** The results can be regarded as a preliminary indication that rents are indeed affected by urban tourism. Rents seem to be positively correlated with the touristic attractiveness of a particular location, even if it is very difficult to accurately measure the real quantity of the respective effects of the urban tourism amenities, as the various models show. *GAM* outperforms the results of *OLS* and seems to be more appropriate for spatial analysis of rents across a city.

**Originality / value:** To the authors' best knowledge, the paper provides the first empirical analysis of the effects of urban tourism hotspots on the Berlin housing market.

**Keywords:** hedonic approach, *GAM*, spatial analysis, tourism hotspots, housing rents, Berlin

## 6.1 Introduction

Berlin has the largest residential rental market in Germany with a ratio of about 85 percent rental housing to owner-occupied and has long had a reputation as a cheap place to live. But recently, soaring rents have threatened that image. Responsible for this development are first and foremost the continuously rising housing demand due to high net inward migration rates (40,000 more inhabitants p.a.) and the limited and comparatively inelastic housing supply (Investitionsbank Berlin, 2013; Berlin Hyp, 2015). Rents have increased particularly in the attractive inner-city districts, but also in the outskirts where demand was comparatively low before. Building permissions as well as building completions have clearly increased and there is confidence that newly constructed flats will ease the situation. In addition, many measures, such as rent control or the misuse-prohibition-law have been introduced to restrict the tightness of the Berlin housing market. However, a significant easing of the housing market can only be achieved, if the additional demand can be satisfied (InvestitionsbankBerlin, 2013).

In addition to high inward migration and a limited housing supply as the main rent booster, urban tourism is held liable for exacerbating the tight situation on the Berlin housing market. After reunification, tourism developed rapidly and today, Berlin is a major tourism destination and one of the three most important players in European urban tourism, beside Paris and London. Arrivals and overnight stays reached 12.6 million (+4.2 percent y-o-y) and 30.3 million (+5.4 percent y-o-y) respectively in 2015, and thus the number of overnight stays grew by leaps and bounds, more than tripling within 16 years. Regarding growth rates, Berlin is even ahead of Paris and London. There were about 240.500 tourism-related jobs, and urban tourism generated €10.65 billion gross revenues, meaning a contribution of 7 percent to aggregate income, and €1.91 billion in taxes for the Federal Government and States in 2014 (VisitBerlin, 2013; Statistical Office Berlin Brandenburg, 2016). That urban tourism has become a truly important economic factor for several cities worldwide is beyond dispute (World Tourism Organization, 2012), whereas the fact that externalities of tourism add value to housing markets has been confirmed only in a few studies (e.g. Biagi and Faggian, 2004; Biagi et al., 2015). In Berlin, many residents are complaining about the massive inbound tourism flows. The local media are full of articles about housing scarcity, displacement and particularly increasing rents due to tourism. In addition, locals are complaining about crowded, rubbish-strewn and noisy neighborhoods. The focus of attention is that besides conventional overnight tourism, a further unrecorded number of tourists stay in private short-term rentals and thus worsen the already tense situation on the housing market, as numerous residential flats are illegitimately misused as short-term holiday flats. Politicans – among others – argue that these short-term rentals additionally stress the housing market. Especially the situation in the inner-city housing districts is tense, where in addition, nearly the entire tourism-infrastructure and a large part of the amenities are located (Schäfer and Braun, 2016; Füller and Michel, 2014).

The aim of this study is to contribute to the literature by providing an approach to quantifying the impact of urban tourism on the housing market, by measuring the spatial effect of tourism hotspots on housing rents and thus contributing to the current debate on ‘touristification’ in Berlin. To the best of the authors` knowledge, there have been no studies in the literature to date, investigating whether urban tourism hotspots affect housing rents. Therefore, the following two null hypotheses were tested:

$H0_{1:}$ Urban tourism hotspots do not affect the housing rents of the surrounding rental flats.

$H0_{2:}$ Different types of urban tourism hotspots cause the same effects.

The research considers *tourist attractions*, *restaurants*, *hotels* and *holiday flats* as constituents of tourism hotspots and is based on a spatial analysis using geographic information systems (*GIS*). The analysis applies a *generalized additive model* (*GAM*) by benchmarking it with conventional *ordinary least squares* (*OLS*), in order to analyze the respective effects on the housing rent.

The paper is structured as follows: Section 6.2 reviews the existing and related literature. Section 6.3 describes the data for the study, and Section 6.4 provides the methodology used, giving a description of the spatial analysis and the hedonic price models. Section 6.5 describes and explains the results of the empirical analysis and Section 6.6 summarizes and concludes the study.

## 6.2 Literature review

Studies on tourism and housing markets are relatively sparse, particularly concerning urban destinations. There is no study measuring the effect of a bundle of urban tourism amenities on the housing rents, while simultaneously considering the attractiveness of each amenity. However, previous research confirms that tourism can certainly affect housing prices. The study of Biagi and Faggian (2004) develops two tourism indices and measures, with an ordinary hedonic pricing model, its effects on Sardinian housing prices at municipality level, and finds a correlation between house prices and tourism, concluding that the more touristic a place, the higher the competitive pressure on the housing market. A further study of Biagi et al. (2015) confirms for 103 Italian cities, that tourism activity affects house prices positively. Most other previous studies analyze the effect of single, predominantly recreation or leisure amenities on housing prices, such as the effect of nearby coasts and beaches (Milon et al., 1984; Pompe and Rinehart, 1995; Rush and Bruggink, 2000; Hamilton, 2007; Conroy and Milosh, 2009), of a certain view (Gillard, 1981; Benson et al., 1998) or of nearby golf courses (Do and Grudnitski, 1995; Nicholls and Crompton, 2007).

Studies on tourism and housing markets regarding Germany and in particular Berlin, are also rare. To the authors' best knowledge there are just two studies: Füller and Michel (2014) describe short-term rentals as a new form of urban tourism in Berlin Kreuzberg and discuss the topic with respect to neighborhood change and gentrification. They find that buying small flats as second homes and/or for short-term rental revenue has become a very attractive investment scheme in Berlin. Accordingly, competition for this highly demanded segment of the housing market continues to increase and really puts pressure on the housing market. Schäfer and Braun (2016) go one step further and analyze the Berlin housing market regarding illegitimate misuse through short-term rentals and find that a significant number of flats are misused as short-term rentals and most of them are centrally located. Furthermore, they find that rental growth is higher in neighborhoods with high numbers of misused flats and conclude that this development contributes to additional pricing pressure on the Berlin housing market.

The present paper also contributes to the limited international literature on the housing market of Berlin. Besides Füller and Michel (2014) and Schäfer and Braun (2016), the following two papers deal with the Berlin housing market from an economic perspective. Ahlfeldt and Maennig (2010) analyze the external effects of heritage-listed buildings on condominium transaction prices in Berlin and find

that such buildings have a positive external effect on surrounding property prices. Kholodilin and Mense (2012) construct rent and price indices for Berlin districts based on their analysis of internet ads.

Most previous studies regarding the measurement of external effects on housing prices or rents applied conventional hedonic pricing models with ordinary least squares (*OLS*) (e.g. Rosen, 1974; Palmquist, 1980). These parametric estimators are highly efficient for well-specified models. However, specification is often very uncertain and the literature reveals that traditional *OLS* methods are often inappropriate, since real estate data can have spatially varying and non-linear effects over space (Tu and Xia, 2008; Hanink et al., 2010; Sunding and Swoboda, 2010; Pace, 1998; Bin, 2004; Cajias, 2014 etc.). As a consequence, several methods have been introduced in recent years to consider these spatially varying and non-linear effects, and hence provide more accurate estimation results regarding real estate prices and rents. A *generalized additive model* (*GAM*) is one of these innovative methods. *GAM* was introduced by Hastie and Tibshirani (1990) and replaces the linear function of conventional *OLS* by a sum of smooth functions, which are more flexible and can control for non-linear relationships. Bin (2004) finds that such a semi-parametric regression with a hedonic price function provides more accurate housing price predictions by decreasing prediction error. *GAM* thus outperforms conventional parametric counterparts and indicates that semi-parametric models would be a more useful tool for the measurement and prediction of housing prices. Cajias (2014) also finds that conventional *OLS* methods are outperformed by *GAM* and another method called *geographic weighted regression* (*GWR*). The results show that *OLS* methods fail to explain housing rents satisfactorily, due to skewed errors and furthermore, that housing rents across 14 German cities respond to spatial and non-linear effects. *GWR* is another method for considering spatially varying effects over space (e.g. Sunding and Swoboda, 2010; Hanink, et al., 2012; Lu et al., 2011). These non-stationary effects enable a more realistic modeling of the housing market with its high degree of spatial heterogeneity. However, *GAM* is more appropriate than *GWR*, since it is more customizable in articulating stationary and non-stationary coefficients. In addition, it operates well with a large sample and simplifies investigating non-linearity (Geniaux and Napoléone, 2008). Consequently, this study forgoes *GWR* and focuses on a hedonic estimation with *GAM*, by benchmarking it with *OLS*.

## 6.3 Data

The analysis comprises the entire city of Berlin, with an area of 892sqkm and 3.47 million inhabitants (as of 2014). Two main data sources are used in this analysis.

The first main dataset is obtained from *Immobilienscout24*, the leading real estate portal in Germany. The provider delivered the quoted housing rents from January 2011 until June 2014. The dataset comprises 511,478 georeferenced raw data (with known longitude and latitude) including structural characteristics (e.g. size, age, rooms etc.) of rental flats in Berlin. After data adjustment and the exclusion of observations with missing values, the analysis was conducted on a sample of 68,984 quoted housing rents.

The second main dataset was collected from *tripadvisor.com*, one of the leading touristic web-sites, which provides user-generated experience reports and ratings for tourism hotspots world-wide. The data on Berlin tourism-hotspots as a whole is freely available on the website of *TripAdvisor*. In total,

7,401 places of tourism (955 tourist attractions, 4,892 restaurants, 666 hotels and 888 holiday flats) with corresponding addresses and ratings as listed in July 2015 were used in this study. The rating of each tourism location describes the attractiveness of each, as perceived by the tourists on a scale from one to five (+1=bad; +5=very good). For the analysis, the corresponding value of attractiveness was determined by combining the quantity of the ratings and the average attractiveness value. These attractiveness scores were then used to derive the final regression variables for each flat by applying two different methods – kernel density estimation and the k-Nearest Neighbor approach (cf. section 6.4.1). The values were standardized in order to make the resulting regression coefficients more comparable.

In addition to these two main datasets, infrastructure and socioeconomic data was collected from various sources. Data on infrastructure (airport, train stations, bus stops), rivers and amenities (supermarkets, pharmacies, kindergartens etc.) was obtained from the open data platform *OpenStreetMap*. The distance to these georeferenced objects was calculated for each rental flat by using GIS. Socioeconomic data (population density, income, unemployment rate and residential area quality) was used according to the 447 planning areas of *LOR*[17]-level 3 in Berlin (Statistical Office Berlin Brandenburg, 2014). Based on their respective address, rental flats were allocated to one of these statistical areas.

The following table presents the data with abbreviations, definitions and descriptive statistics and shows that the average quoted housing rent is 7.60€/sqm for an average living space of 70.54sqm with 2.39 rooms and an average building age of 65.71 years during the period from January 2011 until June 2014. The average time a dwelling is quoted on the market (time on market) is about 2.09 months per flat.

## 6.4 Methodology

### 6.4.1 Processing of spatial data on tourism amenities

As the data on tourism amenities comprises single point values with geo-coordinates, and a respective weighting which depends on the quantity and quality of ratings on *TripAdvisor*, a method is required to assign a certain value of *tourist attractions* (/ *restaurants*/ *hotels*/ *holiday flats*) intensity to each rental flat. Two different methods were tested for this assignment. One approach is based on the *k-Nearest Neighbor approach* with different numbers of nearest amenities of each category and is calculated directly for each rental flat. The second approach is based on kernel density estimation and creates a density surface for each amenity category for the entire city. Spatial interpolation techniques like *Kriging* are not suitable for this purpose. *Kriging* also creates a surface of values based on a set of localized input values. However, it only reflects the local parameter values, and not the density of the points themselves. *Kriging* will be applied for evaluating the spatial distribution of residuals of the regression models.

[17] On 01.08.2006 the Berlin Senate passed the *LOR* (life-world orientated areas) as the new spatial base for planning, predicting and monitoring demographic and social developments. This is the smallest statistical unit for providing demographic and socioeconomic population data in Berlin.

**Table 1: Variables, definitions and descriptive statistics**

| Variable | Definition | Mean | St. dev. |
|---|---|---|---|
| *Dependent Variable* | | | |
| R* | Average quoted rent of flat per sqm in € | 7.60 | 2.19 |
| *Property Characteristics (PC)* | | | |
| PC_FS* | Floor space in sqm | 70.54 | 32.46 |
| PC_AD | Age of dwelling in years | 65.71 | 40.78 |
| PC_SR* | Average space per room | 30.75 | 8.33 |
| PC_KN | 1 if property has a kitchen, 0 otherwise | 0.49 | 0.50 |
| PC_BY | 1 if property has a balcony, 0 otherwise | 0.68 | 0.46 |
| PC_ToM* | Time on market of dwelling | 2.09 | 3.87 |
| *Neighborhood (NBH)* | | | |
| NBH_RAQ | Quality of residential area | 1.81 | 0.66 |
| NBH_PD* | Population density | 10,773.59 | 7,230.77 |
| NBH_UR* | Mean unemployment rate in statistical district (in %) | 8.75 | 3.72 |
| NBH_MB | Mean migration balance in statistical district (migrants per 1.000 inhabitants) | 12.35 | 13.83 |
| *Access (AS)* | | | |
| AS_AP* | Distance to next airport (in meters) | 9,264.50 | 3,669.86 |
| AS_TS* | Distance to next train station (in meters) | 758.13 | 688.07 |
| AS_BS* | Distance to next bus stop (in meters) | 235.87 | 236.84 |
| AS_SM* | Distance to next supermarket (in meters) | 386.16 | 319.72 |
| AS_BK* | Distance to next bakery (in meters) | 345.10 | 344.76 |
| AS_PM* | Distance to next pharmacy (in meters) | 372.09 | 351.42 |
| AS_KG* | Distance to next kindergarten (in meters) | 280.04 | 256.37 |
| AS_RV* | Distance to next river (in meters) | 1,029.12 | 781.16 |
| *Urban tourism hotspots (UTH)* | k-Nearest Neighbor approach (kNN) | | |
| UTH_1NN_TA | Score variable for nearest tourist attraction | 0.01 | 0.08 |
| UTH_1NN_RT | Score variable for nearest restaurant | 0.05 | 0.77 |
| UTH_1NN_HL | Score variable for nearest hotel | 0.01 | 0.07 |
| UTH_1NN_HF | Score variable for nearest holiday flat | 0.02 | 0.20 |
| UTH_3NN_TA | Score variable for 3 nearest tourist attractions | 0.01 | 0.08 |
| UTH_3NN_RT | Score variable for 3 nearest restaurants | 0.08 | 0.84 |
| UTH_3NN_HL | Score variable for 3 nearest hotels | 0.02 | 0.08 |
| UTH_3NN_HF | Score variable for 3 nearest holiday flats | 0.04 | 0.31 |
| UTH_5NN_TA | Score variable for 5 nearest tourist attractions | 0.01 | 0.09 |
| UTH_5NN_RT | Score variable for 5 nearest restaurants | 0.09 | 0.84 |
| UTH_5NN_HL | Score variable for 5 nearest hotels | 0.03 | 0.09 |
| UTH_5NN_HF | Score variable for 5 nearest holiday flats | 0.05 | 0.38 |
| *Urban tourism hotspots (UTH)* | Weighted Kernel Density Estimation (wKDE) | | |
| UTH_wKDE_0.5_TA | Density variable (bw=0.5km) for tourist attractions | 25.89 | 72.08 |
| UTH_wKDE_0.5_RT | Density variable (bw=0.5km) for restaurants | 64.67 | 136.59 |
| UTH_wKDE_0.5_HL | Density variable (bw=0.5km) for hotels | 36.71 | 100.07 |
| UTH_wKDE_0.5_HF | Density variable (bw=0.5km) for holiday flats | 20.65 | 49.81 |

Note 1: The variables with * are logarithmized in the regression models below. The descriptive statistics show their natural values.
Note 2: Due to the high correlation of average living space and the number of rooms, the latter variable was replaced by the average size of rooms. At 0.06, the correlation of living space and the average size of rooms is comparatively low and refutes the objection, that larger flats might naturally have larger rooms. The correlation matrix is shown in Appendix 1.

#### 6.4.1.1 k-Nearest Neighbor approach

The *k-Nearest Neighbor* approach is based on a score, which is assigned to each rental flat for each amenity category, by adding up the distance-weighted scores of the *k* (with *k = 1, 3, 5*) nearest representatives of the respective amenity category. First, for each rental flat, the *k* nearest tourist amenities of each category were determined. Below, the procedure for the case of *tourist attractions* is described. The amenity categories *restaurants*, *hotels* and *holiday flats* are treated analogously. A set of *tourist attractions* $TA = \{ta_1, \dots, ta_n\}$ is defined, where *n* is the total number of *tourist*

*attractions* in the analysis. Furthermore, a set of all *N=68,984* rental flats $F = \{f_1, \dots, f_N\}$ is defined. $d(f_i, ta_j)$ denominates the Euclidean distance between rental flat *i* and *tourist attraction j*. The set $kNN_{TA}(f_i)$ of those *k* nearest *tourist attractions* for rental flat *i* can now be defined according to Papadias et al. (2004), as those *k tourist attractions* with the smallest sum of distances to $f_i$:

$$kNN_{TA}(f_i) = \left\{ ta \left| \sum_{ta_j \in TA^*} d(f_i, ta_j) \leq \sum_{ta_k \in TA^*} d(f_i, ta_k);\ j \neq k;\ j = 1, \dots, n; k = 1, \dots, n; TA^* \in TA; |TA^*| = k \right. \right\} \quad (1)$$

*with $|TA^*|$ = Cardinality of set $TA^*$ and k = (1, 3 ,5)*

The *k* elements of $kNN_{TA}(f_i)$ shall be denoted as $kNN_{TA}(f_i)_m$ with $m = 1, \dots, k$.

Each *tourist attraction* $ta_j$ was assigned a score $STA^{kNN}(h_j)$, depending on the number $nTA_j$ and quality $QTA_j$ of ratings on *TripAdvisor*:

$$STA^{kNN}(h_j) = QTA_j * ln(nTA_j + 1) \quad (2)$$

The ratings $QTA_j$ are rescaled to the range from -2.5 to +2.5 (the *TripAdvisor*-ratings originally ranged from +1 to +5). This accounts for the potential negative impacts of amenities with very poor quality and/or reputation. Without this transformation, both high quality amenities with few ratings and low quality amenities with many ratings, would score comparable (low, positive) values. Considering the low number of *holiday flats* with any rating at all, the score in this amenity category is composed of the logarithmized number of beds only.

Finally, each rental flat $f_i$ is assigned an amenity-specific score (in the case of *tourist attractions* denominated as $kNN_{TA}(f_i)$), which depends on the scores of the nearest *k* amenities of the specific category $STA^{kNN}(kNN_{TA}(f_i)_m)$ and their distance $d(f_i, kNN_{TA}(f_i)_m)$ to the flat:

$$kNN_{TA}f_i = \sum_{m=1}^{X} \frac{STA^{kNN}(kNN_{TA}(f_i)_m)}{d(f_i, kNN_{TA}(f_i)_m)} \quad (3)$$

These values are used in the final regression models with *OLS* and *GAM*. The calculations were conducted with the programming language *R* (R Core Team, 2013).

#### 6.4.1.2 Weighted Kernel Density Estimation

In the approach with a *weighted Kernel Density Estimation* (*wKDE*), each rental flat is also assigned a certain value for each amenity category. In contrast to the *k-Nearest Neighbor approach*, the number of representatives of each amenity category is not fixed, but depends on the number of these representatives below a certain threshold around the respective flat.

*Kernel density estimation* (*KDE*) is a non-parametric method, which enables the construction of a continuous probability density function, based on a given set of data points. In the present setting, this density function is a two-dimensional surface which is constructed for each amenity category before each rental flat is assigned the local values of these surfaces. The density surfaces in Figure 1 show distinct patterns for the four analyzed amenity categories, with a bandwidth of 0.5 kilometers. Broadly speaking, in *KDE*, a so-called kernel function is applied to each data point, creating a distance-weighted density curve around it. All of these kernel density functions are added up and the resulting density function is scaled, so that the area (1D) or volume (2D, like in the present case) has the value 1. The well-known *Epanechnikov kernel* was applied, which minimizes the mean squared

error (*MSE*) of the estimation (Epanechnikov, 1969). The *Epanechnikov kernel* is a so-called bounded kernel, which only considers the influence of those data points within the chosen bandwidth. Basically, the selection of this bandwidth influences the smoothness of the *KDE* function or the spatial range in which a data point affects the respective results. If the chosen bandwidth is too small, the surface will be composed of single peaks for each amenity. If the chosen bandwidth is too high, local variations will be blurred. The authors decided not to apply any formal bandwidth selection technique, relying on a value of 0.5 kilometers, which seemed most appropriate for the special case of tourist amenities and the range of their effects. Nevertheless, it should be mentioned that the results might change for a higher or lower bandwidth. [18]

The procedure applied in the present study uses a *wKDE*, which takes into account a defined score of each amenity, when estimating the kernel density. As the weight function has to be non-negative, the ratings of *TripAdvisor* were not rescaled in this approach and range from +1 (very bad) to +5 (very good). The following formula refers to the case of the amenity category *tourist attractions*. The weight of each *tourist attraction* $(ta)$ was calculated by combining the number $nTA_j$ and quality $QTA_j$ of ratings on *TripAdvisor*:

$$STA^{wKDE}(ta_j) = (QTA_j + 1) * (ln(nTA_j + 1) + 1) \quad (4)$$

The addition of one to the number of ratings and the logarithmized quality of the rating ensures that the minimum weight is one for *tourist attractions* without any rating. The quality $QTA_j$ was defined as zero in this case. The addition of one to the number of ratings accounts for the fact that the logarithmic function is defined only for numbers greater than zero.

The value of the *wKDE* at a specific location $x$ is defined as follows. $ta_1, \dots, ta_j \in \mathbb{R}^2$ be a set of *tourist attractions*, $K$ a kernel, $w_j$ the weighting factor with $\sum_{j=1}^{n} w_j = 1$ and $h > 0$ the bandwidth.

$$f_h(x) = \frac{1}{h}\sum_{j=1}^{n} w_j \, K\left(\frac{x - x_j}{h}\right) \quad (5)$$

with $x_j$ = location of *tourist attraction* $j$; *Epanechnikov kernel* $K(x) = \begin{cases} 3/4\,(1-x^2), & |x| \le 1 \\ 0, & |x| > 1 \end{cases}$

### 6.4.2 Hedonic approach

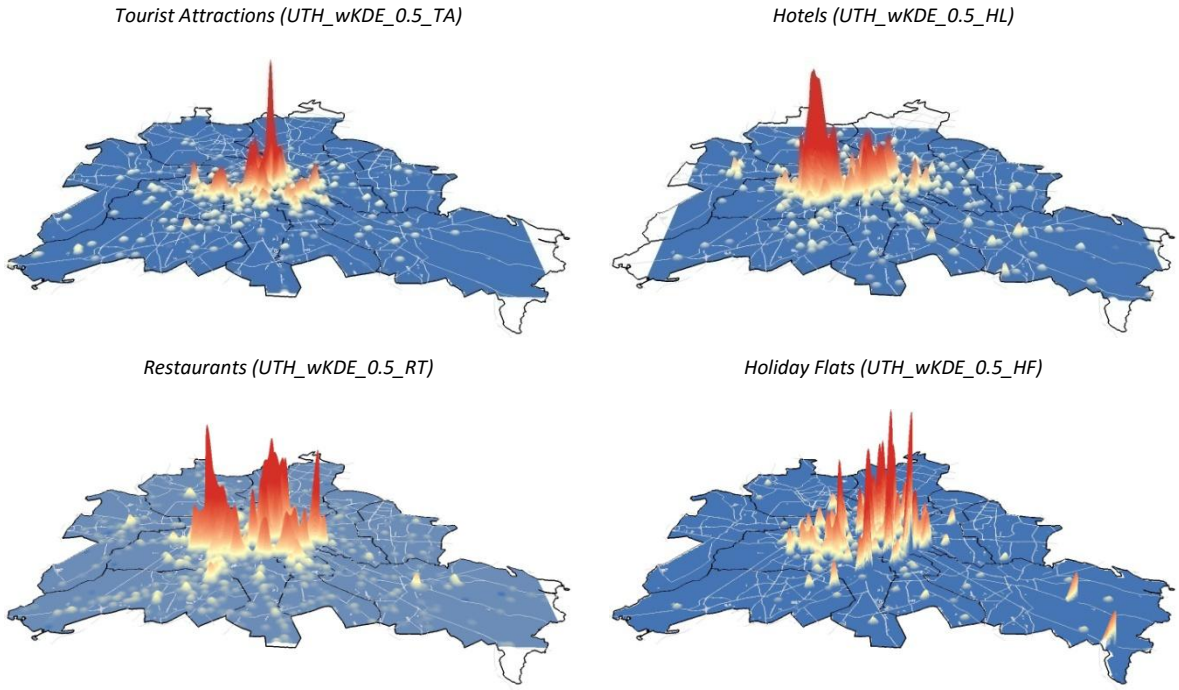
Conventional hedonic approaches explain the price of a composite good by a bundle of explanatory variables using *ordinary least squares* (*OLS*). The basic parametric *OLS* model can be written as:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i \mathrm{X}_i + \varepsilon \quad (6)$$

Where $Y$ is the response variable, $\beta_0$ is the intercept, $p$ is the number of predictor variables, $\beta_i$ is the estimation parameter for observation $i$, $X_i$ is the predictor variable for observation $i$ and $\varepsilon$ is the random error (Baranzi et al., 2008). In terms of real estate, the hedonic approach explains the price or the rent of a dwelling by its *structural attributes* such as floor space and building age, by its

[18] Further models with a bandwidth of bw=1.0km and bw=2.5km were rejected due to strong multicollinearity.

**Figure 1: Weighted Kernel Density Estimation for all amenity categories across Berlin with a bandwidth of 0.5 kilometers**




Note: The plots show where the respective density of the attractive *UTH* is high/low. The dark-red piles illustrate the most attractive places on the Berlin map. Obviously, *UTH* are located very centrally.

*neighborhood characteristics* such as population density and unemployment rate and by its *locational characteristics* such as its access to public transportation services or grocery stores (Rosen, 1974; Zheng and Kahn, 2008).

This straightforward and linear functional form is often criticized for not being able to convey sufficient information to enable a correct and successful specification of parametric models. If the functional form is not precise or further important variables are omitted, the results are highly questionable. Another drawback is that the consideration of spatial variation across space is restricted. However, it is possible to include submarket binary variables to capture such spatial effects as city districts (Bao and Wan, 2004; Cajias, 2014*).*

Due to enhanced computer technology and statistical software on parametric regression, more flexible approaches have consequently received more attention and recognition in the last few years. These models are more useful and more flexible for estimating the price of a dwelling. One of these approaches is a so-called generalized additive model (GAM). This is an extension of the *generalized lineal model* (*GLM*) and extends the linear form $\sum \beta_i X_i$ by a sum of non-linear smoothing functions $\sum f_i (X_i)$. Semi-parametric models are estimated by the iterative procedure, the so-called '*back fitting algorithm'* (Hastie and Tibshirani, 1990). The conventional model can be written as:

$$Y = f_0 + \sum_{i=1}^{p} f_i(X_i) + \varepsilon \qquad (7)$$

where the errors $\varepsilon$ are independent of the $X_i$´s, $E(\varepsilon) = 0$ and $Var(u) = \sigma^2$. The $f_i$´s are arbitrary univariate functions, one for each variable.

With *GAM*, spatial variation and non-linear effects can be considered simultaneously. Non-linear effects of rents, regarding for instance the size of a dwelling, can thus be taken into consideration. If a (commonly two-dimensional) spatial variable is incorporated, the model is referred to as '*geoadditive'*. It is essential to choose the right smooth function to capture the existing non-linear effects. Smooth functions or smoothing splines are essentially piecewise polynomials whose different polynomial segments are tied together in a series of knots in a way that ensures certain continuity properties (Eubank, 1999). The following analysis applies *P-splines* and *thin-plate regression splines*. *P-splines* are based on normal *B-splines* and incorporate a so-called penalty-function, which allows for a high number of knots (and thus flexibility), but avoids overfitting by penalizing differences of adjacent parameters (Eilers and Marx, 1996). The smoothing parameter $\lambda$ controls the degree of smoothing. The *MGCV* package for *R* offers a range of *GAM*-related functions. Using this package, the optimal smoothing parameter $\lambda$ for the splines is determined by minimizing the *generalized cross-validation* (*GCV*) score.

For a given $\lambda$, the GCV score is

$$GCV(\lambda) = \frac{\frac{1}{n}\|(I - A(\lambda))y\|^2}{\left[\frac{1}{n}tr(I - A(\lambda))\right]^2} \qquad (8)$$

with $I$ the identity matrix and $A(\lambda)$ the hat matrix for a given $\lambda$ (Craven and Wahba, 1979).

*Thin-plate regression splines* (*TPRS*) were applied to the spatial variable defined by the geo-coordinates of the flats and to the term representing the interaction between location and the age of the flats. *TPRS* also penalize overfitting and represent a more flexible and computationally efficient version of standard *thin-plate splines* (Wood, 2003; Duchon, 1976). The name is derived from their physical motivation to minimize the bending energy of a thin metal plate (Bookstein, 1989). *TPRS* are particularly suitable for the modeling of two-dimensional variables like the location, by avoiding the problems of knot placement (Wood, 2003).

### 6.4.3 Model specification

Based on the above theory the adjusted *OLS* and *GAM* for the following hedonic analysis is as follows:

*OLS*

$$ln\,R = \beta_0 + \beta_1 PC + \beta_2 NBH + \beta_3 AS + \beta_4 TEY + \beta_5 SED + (\beta_6 UTH) + \varepsilon \qquad (9)$$

where *lnR* is a vector of the natural logarithm of the quoted housing rents, $\beta_0$ is the intercept, $\beta_1$-$\beta_6$ are the vectors of coefficients of the estimated parameters, *PC* is a matrix of the flats defining physical characteristics, *NBH* is a matrix of neighborhood characteristics, AS is a matrix of the flats defining access characteristics, *TEY* is a vector of time components, *SED* is a matrix which controls for spatial effects between the single city districts, *UTH* is a matrix of the flats defining tourism characteristics and$\varepsilon$is a vector of random errors. The parentheses denote that *UTH* are not included in every model, in order to be able to measure the marginal effect of *UTH*. The same holds for equation 10.

*GAM*

$$ln\,R = \beta_0 + \beta_1 PC_B + s_1 PC_S + s_2 NBH + s_3 AS + \beta_2 TEY + s_4 SELA(u_i, v_i) + s_5 UTH + \varepsilon \qquad (10)$$

where *lnR* is a vector of the natural logarithm of the quoted housing rents, $\beta_0$ is the intercept, $\beta_1$ and $\beta_2$ are the vectors of coefficients of the respective estimated parameters, $PC_B$ is a matrix of the flats defining physical characteristics (in this case, binary variables for kitchen and balcony) and *TEY* is a vector for the time component Year. $s_1$-$s_5$ are smooth functions, $PC_S$ is a matrix of the remaining non-dichotomous flats defining physical characteristics (splines), *NBH* is a matrix of the flats defining neighborhood characteristics, *AS* is a matrix of the flats defining access characteristics, *SELA* is a matrix which controls for further spatial effects regarding location and its interaction with building age, *UTH* is a matrix of the flats defining tourism characteristics and $\varepsilon$ is a vector of random errors.

## 6.5 Empirical results

Based on the two approaches above, ten models were estimated in total, five (M1-M5) with *OLS* and five (M6-M10) with *GAM*. One in each case without *UTH*, three with *UTH* using the *k-Nearest Neighbor approach* and one with *UTH* using *weighted Kernel Density Estimation* (cf. section 6.4.1). The principal objective of this study is to measure whether *UTH* affect housing rents. The results for the ten models are presented below:

### 6.5.1 OLS (M1-M5)

**General control results:** Table 2 represents the regression results for *OLS*. Columns I-V show the ordinary least squares (*OLS*) without *UTH* (M1), with *UTH* using the *k-Nearest Neighbor approach* (M2-M4) and with *UTH* using *weighted Kernel Density Estimation* (M5).

Almost all variables (except *NBH_MB*) are highly significant and provide meaningful results with signs that are in line with those in the literature. Only a few signs are not in line with general expectations and are presumably attributable to non-linear relationships. For instance, the property characteristic *PC_FS* is generally expected to be positive. However, if there are, for instance, many small flats which are comparatively more expensive than larger ones, it might be that the sign changes. The other property characteristic variables are in line with the generally expected sign and the variable *PC_ToM* also confirms the expectations that the longer a rental flat is on the market, ceteris paribus, the higher its rent (cf. Knight, 2008). The signs of the access variables are also not straightforward. For instance, a good infrastructure with, for instance, good public transportation service and supermarkets, is expected to generally add value to the surrounding housing market, but on the other hand, the proximity to a train station or supermarket should not be too close. It is assumed that a non-linear consideration would be more sensible in the cases mentioned. This approach aims to explain a few generally unexpected signs and additionally shows that the linearity of OLS models can reflect reality only to a certain extent. The neighborhood variables are in line with the generally expected sign. Controlling for spatial effects, the district coefficients are distinctly different across the single districts. Friedrichshain-Kreuzberg *(SED_FK, +29 percent)*, Charlottenburg-Wilmersdorf *(SED_CW, +26 percent)*, Mitte *(SED_MT, +26 percent)* and Pankow *(SED_PW, +21 percent)* have the highest positive correlations with the rents, and hence the priciest rental markets in Berlin. [19] When considering the time effect (*TEY*) of the data, all five models show a significant increase of approximately 7.0 percent p.a. This is slightly lower than the actual official numbers, indicating that some share of the average increase can be explained by improved building characteristics, but the

[19] The district 'Marzahn-Hellersdorf' represents the reference category and is accordingly not listed. The same holds for the time effect variable of 2011.

major part remains market-driven. The time effect (*TEY*) can be further affected by an increased number of flats rented out through *Immobilienscout24* in high-rent districts. Actually, an approximately proportional relationship between the OLS-coefficient of a district and the relative change in the number of rented flats was found, indicating a growing interest of investors in these high-price areas. The introduction of an interaction term between district and year (*TEY*) reveals that the impact of the described change of spatial priorities on the general time effect is rather small, as the coefficient of *TEY* still amounts to 6.0 percent. The interaction model further reveals large and significant rent increases in every year for the districts Mitte, Friedrichshain-Kreuzberg and Neukölln, which are of considerable interest to investors at present.

**Effects of UTH on housing rents:** The main aim of this analysis is to analyze whether urban tourism affects housing rents. By adding the respective variables of *UTH* (*tourist attractions, restaurants, hotels and holiday flats*), all four models M2-M5 improve, even if only slightly and differently, regarding the explained deviation and minimizing *MSE*, and in particular show that the single *UTH* affect housing rents. The results depend on the type of the model and its respective applied approach. M4 (*5NN*), for instance, shows that an increase in the *tourist attraction* variable *(UTH_5NN_TA)* of one standard deviation leads to an increase in the housing rents of 0.84 percent, the increase in the *restaurant* variable *(UTH_5NN_RT)* leads to an increase in the housing rents of 0.42 percent, the increase of the *hotel* variable *(UTH_5NN_HL)* leads to an increase in the housing rents of 1.35 percent and the increase of the *holiday flat* variable *(UTH_5NN_HF)* to an increase in the housing rents of 0.38 percent by adding one unit, ceteris paribus. The results change, when considering M5 (*wKDE* with bw=0.5km). *Tourist attractions (UTH_wKDE_0.5_TA)* increase housing rents by 0.96 percent, *restaurants (UTH_wKDE_0.5_RT)* by 2.46 percent, hotels *(UTH_wKDE_0.5_HL)* by 1.00 percent and *holiday flats (UTH_wKDE_0.5_HF)* by 1.42 percent by adding one unit, ceteris paribus. However, there is a moderate correlation of 0.58 between *tourist attractions* and *restaurants* (cf. Appendix 1) and this indicates multicollinearity. Hence, the results for M5 are presumably biased and thus questionable to a certain extent. Nevertheless, when considering the *UTH_wKDE* variables separated from each other in further models, the results of these unbiased models confirm the positive effects of the single *UTH* categories on the housing rents.

Here, *tourist attractions (UTH_wKDE_0.5_TA_S)* show an increase of 0.42 percent on the surrounding housing rents. *Restaurants* (*UTH_wKDE_0.5_RT_S*) show an increase of 1.21 percent, *hotels* (*UTH_wKDE_0.5_HL_S*) of 1.82 percent and *holiday flats* (*UTH_wKDE_0.5_HF_S*) of 2.56 percent on the housing rents within a bandwidth of 0.5 kilometers.

In summary, the results of the four *OLS* models which consider *UTH* variables support the current debate that tourism is likely to add value to housing rents and also show that it is very difficult to measure the correct and individual effect from the single tourist amenities, and that the single *UTH* variables affect the housing rents differently.

The next section provides the results of the semi-parametric approach by considering non-linear effects.

**Table 2: Results of OLS (M1-M5)**

| | OLS | | | | |
|---|---|---|---|---|---|
| | **M1 without UTH** | **M2 with UTH 1NN** | **M3 with UTH 3NN** | **M4 with UTH 5NN** | **M5 with UTH wKDE_0.5** |
| Intercept | 1.657*** | 1.6610*** | 1.6710*** | 1.680*** | 1.782*** |
| *Property Characteristics (PC)* | | | | | |
| PC_FS | -0.0623*** | -0.0625*** | -0.0629*** | -0.0632*** | -0.0670*** |
| PC_AD | 0.0003*** | 0.0003*** | 0.0003*** | 0.0003*** | 0.0003*** |
| PC_SR | 0.0532*** | 0.0526*** | 0.0519*** | 0.0514*** | 0.0454*** |
| PC_KN[1] | 0.0905*** | 0.0903*** | 0.0900*** | 0.0898*** | 0.0874*** |
| PC_BY[1] | 0.0203*** | 0.0204*** | 0.0206*** | 0.0207*** | 0.0223*** |
| PC_ToM | 0.0148*** | 0.0147*** | 0.0147*** | 0.0146*** | 0.0145*** |
| *Neighborhood (NBH)* | | | | | |
| NBH_RAQ | 0.0172*** | 0.0170*** | 0.0162*** | 0.0155*** | 0.0154*** |
| NBH_PD | 0.0344*** | 0.0339*** | 0.0331*** | 0.0323*** | 0.0212*** |
| NBH_UR | -0.2133*** | -0.2118*** | -0.2098*** | -0.2080*** | -0.1836*** |
| NBH_MB | 0.0001* | 0.0001* | 0.0001* | 0.0002* | 0.0001 |
| *Access (AS)* | | | | | |
| AS_AP | 0.0454*** | 0.0448*** | 0.0437*** | 0.0428*** | 0.0344*** |
| AS_TS | -0.0325*** | -0.0321*** | -0.0312*** | -0.0305*** | -0.0251*** |
| AS_BS | 0.0276*** | 0.0275*** | 0.0274*** | 0.0273*** | 0.0231*** |
| AS_SM | -0.0089*** | -0.0086*** | -0.0082*** | -0.0079*** | -0.0046*** |
| AS_BK | -0.0119*** | -0.0116*** | -0.0113*** | -0.0111*** | -0.0077*** |
| AS_PM | 0.0057*** | 0.0059*** | 0.0061*** | 0.0062*** | 0.0097*** |
| AS_KG | -0.0071*** | -0.0072*** | -0.0074*** | -0.0076*** | -0.0084*** |
| AS_RV | -0.0119*** | -0.0118*** | -0.0115*** | -0.0113*** | -0.0111*** |
| *Spatial Effects Districts (SED)* | | | | | |
| SED_CW | 0.2580*** | 0.2567*** | 0.2540*** | 0.2516*** | 0.2395*** |
| SED_FK | 0.2878*** | 0.2856*** | 0.2811*** | 0.2771*** | 0.2557*** |
| SED_LB | 0.0977*** | 0.0976*** | 0.0972*** | 0.0968*** | 0.1070*** |
| SED_MT | 0.2576*** | 0.2546*** | 0.2485*** | 0.2432*** | 0.2155*** |
| SED_NK | 0.1557*** | 0.1553*** | 0.1538*** | 0.1523*** | 0.1543*** |
| SED_PW | 0.2095*** | 0.2086*** | 0.2069*** | 0.2055*** | 0.1945*** |
| SED_RD | 0.1042*** | 0.1038*** | 0.1027*** | 0.1017*** | 0.0976*** |
| SED_SD | 0.1080*** | 0.1067*** | 0.1039*** | 0.1014*** | 0.0892*** |
| SED_SZ | 0.1241*** | 0.1251*** | 0.1266*** | 0.1276*** | 0.1445*** |
| SED_TS | 0.1585*** | 0.1585*** | 0.1581*** | 0.1575*** | 0.1587*** |
| SED_TK | 0.1310*** | 0.1303*** | 0.1285*** | 0.1270*** | 0.1217*** |
| *Time Effects Years (TEY)* | | | | | |
| TEY_2012 | 0.0729*** | 0.0729*** | 0.0729*** | 0.0730*** | 0.0732*** |
| TEY_2013 | 0.1469*** | 0.1466*** | 0.1465*** | 0.1464*** | 0.1464*** |
| TEY_2014 | 0.2102*** | 0.2099*** | 0.2098*** | 0.2097*** | 0.2093*** |
| *Urban Tourism Hotspots (UTH)* | | | | | |
| UTH_1NN_TA | - | 0.0046*** | - | - | - |
| UTH_1NN_RT | - | 0.0030*** | - | - | - |
| UTH_1NN_HL | - | 0.0056*** | - | - | - |
| UTH_1NN_HF | - | 0.0032*** | - | - | - |
| UTH_3NN_TA | - | - | 0.0069*** | - | - |
| UTH_3NN_RT | - | - | 0.0036*** | - | - |
| UTH_3NN_HL | - | - | 0.0106*** | - | - |
| UTH_3NN_HF | - | - | 0.0038*** | - | - |
| UTH_5NN_TA | - | - | - | 0.0084*** | - |
| UTH_5NN_RT | - | - | - | 0.0042*** | - |
| UTH_5NN_HL | - | - | - | 0.0135*** | - |
| UTH_5NN_HF | - | - | - | 0.0038*** | - |
| UTH_wKDE_0.5km_TA | - | - | - | - | 0.0096*** |
| UTH_wKDE_0.5km_RT | - | - | - | - | 0.0246*** |
| UTH_wKDE_0.5km_HL | - | - | - | - | 0.0100*** |
| UTH_wKDE_0.5km_HF | - | - | - | - | 0.0142*** |
| **N** | **68,984** | **68,984** | **68,984** | **68,984** | **68,984** |
| **Adj. $R^2$ (%)** | **45.8** | **45.9** | **46.1** | **46.2** | **46.9** |
| **MSE** | **0.040** | **0.040** | **0.040** | **0.040** | **0.039** |

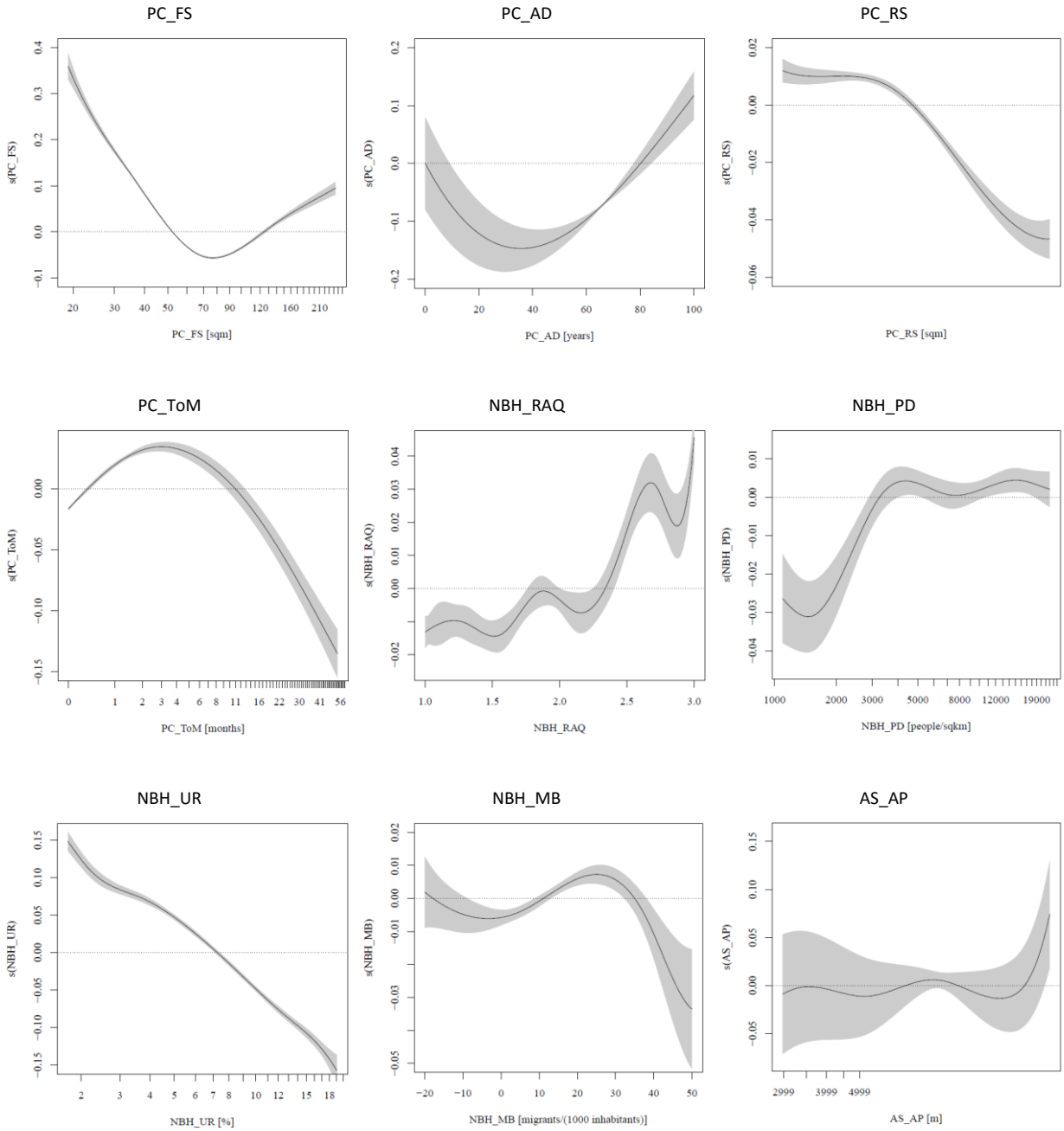
* (0.05) ** (0.01) *** (0.001); [1]= Binary Variables

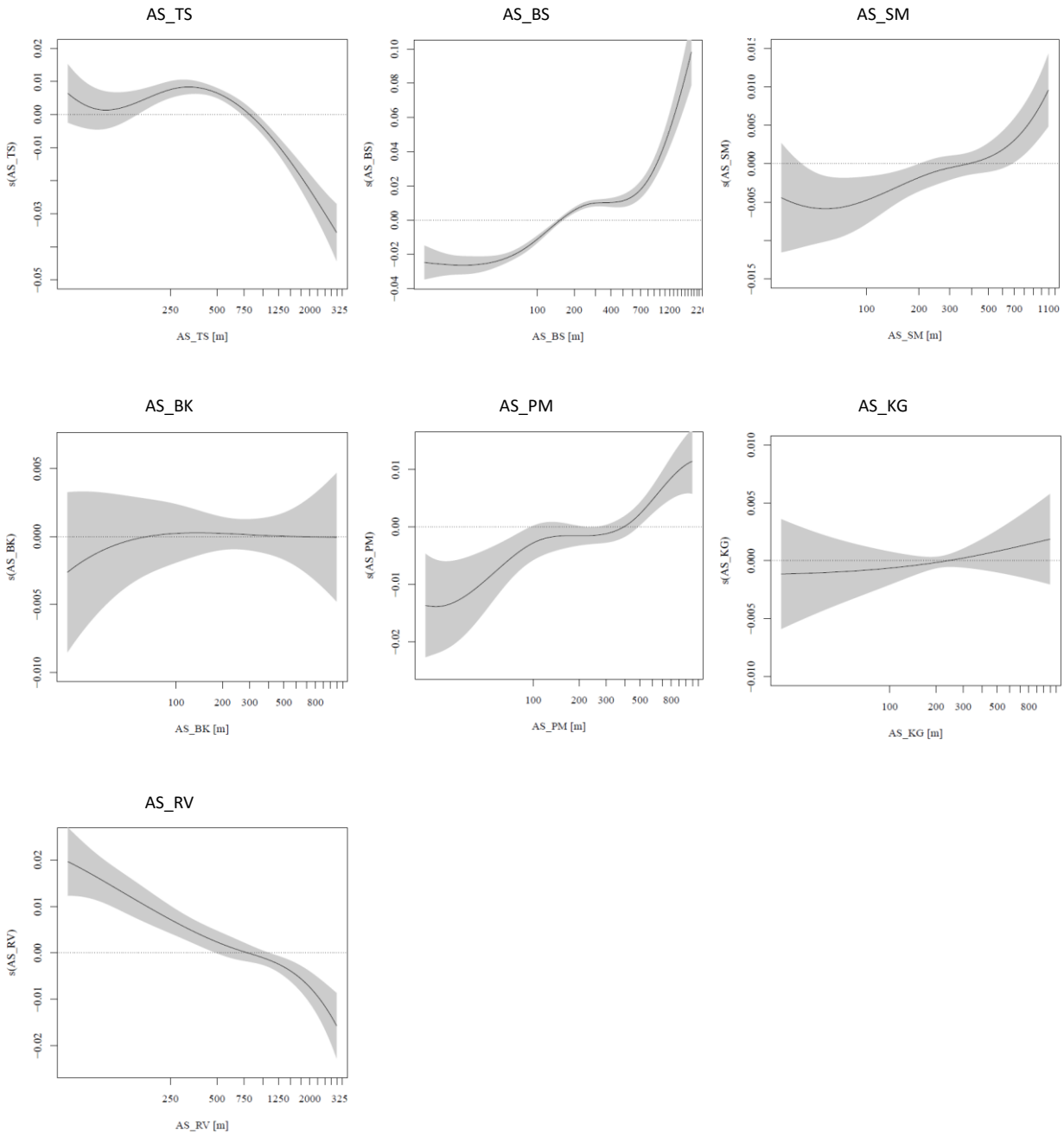
Note: The dependent variable is the logarithmized housing rent/sqm.

### 6.5.2 GAM (M6-M10)

**General control results:** *GAM* accounts for non-linear effects. The illustration of the splines best shows these non-linear effects of the single hedonic characteristics. Only *kitchen*, *balcony* and *year* were considered as binary variables in the five models M6-M10. Their estimates are not presented here, but have the same sign as the *OLS* estimates and are highly significant. The magnitudes of these coefficient estimates are also comparable to the *OLS* models. The general control results of the non-dichotomous variables are very similar for M6-M10 and do not deviate significantly from one another. For illustration purposes, only the results of M9 are presented in the following figure. All estimates are highly significant.

**Figure 2: Effects of hedonic characteristics on housing rents (M9)**



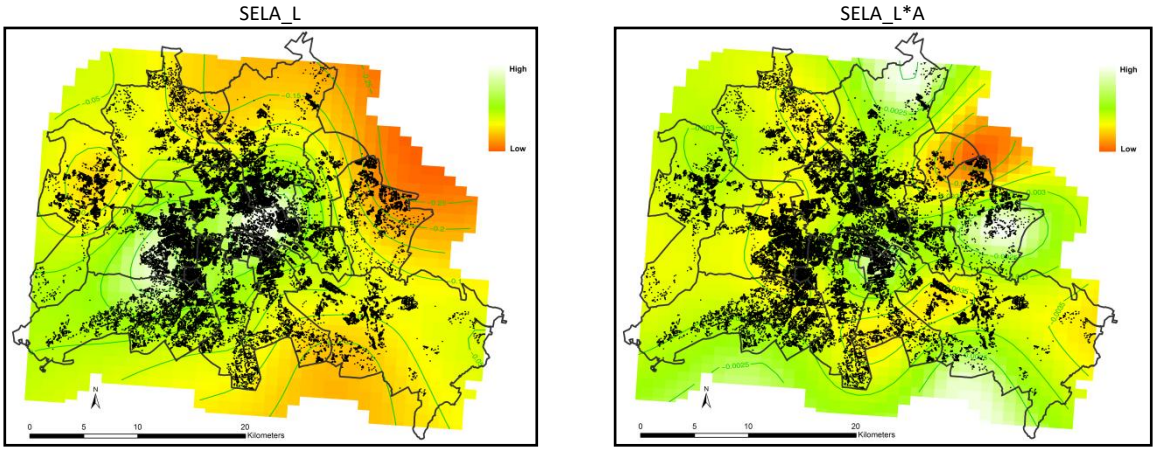



Note: The solid line represents the semi-parametric regression estimates, and the gray shaded area the 95 percent semi-parametric confidence interval estimates.

At first glance, the estimates of hedonic characteristics clearly show the advantage of *GAM* compared to *OLS*. The parametric functional form of *OLS* seems to be inappropriate for accurately approximating the complex effects on housing rents of most of the variables. For instance, the estimated regression of *PC_FS* depicts an initial negative effect, meaning the smaller the dwelling, the higher its rent. This effect changes with a living space of approximately 70-80sqm. These dwelling sizes (and larger) show a positive effect on housing rents, before the effect disappears with large dwellings with more than approximately 200sqm, showing that additional units do not add significant value. This trend may perhaps explain to a certain extent the negative sign of *PC_FS* in the *OLS*-models.

The estimated regression for *PC_AD* depicts a convex relationship, meaning that the housing rent decreases with increasing age for relatively young dwellings. After the dwelling is about 40 years old, the effect disappears and indicates increasing value for well modernized and renovated dwellings of the *Gründerzeit*[20]. *NBH_RAQ* depicts a fitted function with an upward-trend showing clearly that the better the quality of the residential area, e.g. measured by infrastructure quality, the higher the respective dwelling rent. *AS_BS* depicts a fitted upward-slope function, and indicates that proximity to a bus stop is favorable. However, the very close proximity with a distance of up to 500 meters adds no value, possibly due to close and noisy streets or noisy passengers waiting for public transport. The positive effect on rents (between about 500 and 2,000 meters) with an appropriate safe distance indicates the access to good public transportation services. In total, the results show that *OLS* is not able to consider these complex non-linear relationships of responses and predictors. The remaining regression estimates for the *P-splines* can also be interpreted in this way. Lastly, in order to consider further disregarded spatial effects, the model controls for *location* (*SELA_L*) and *location*age* (*SELA_L*A*). The spatial effects were estimated using *thin-plate regression splines* to more effectively show their spatial distribution across Berlin. The results for M9 can be seen in the following figure.

**Figure 3: Effect of spatial parameters (location and location*age) on housing rents (M9)**




Note: Black dots represent the analyzed residential flats, dark-grey lines represent district borders, green lines represent contour lines of the respective coefficient.

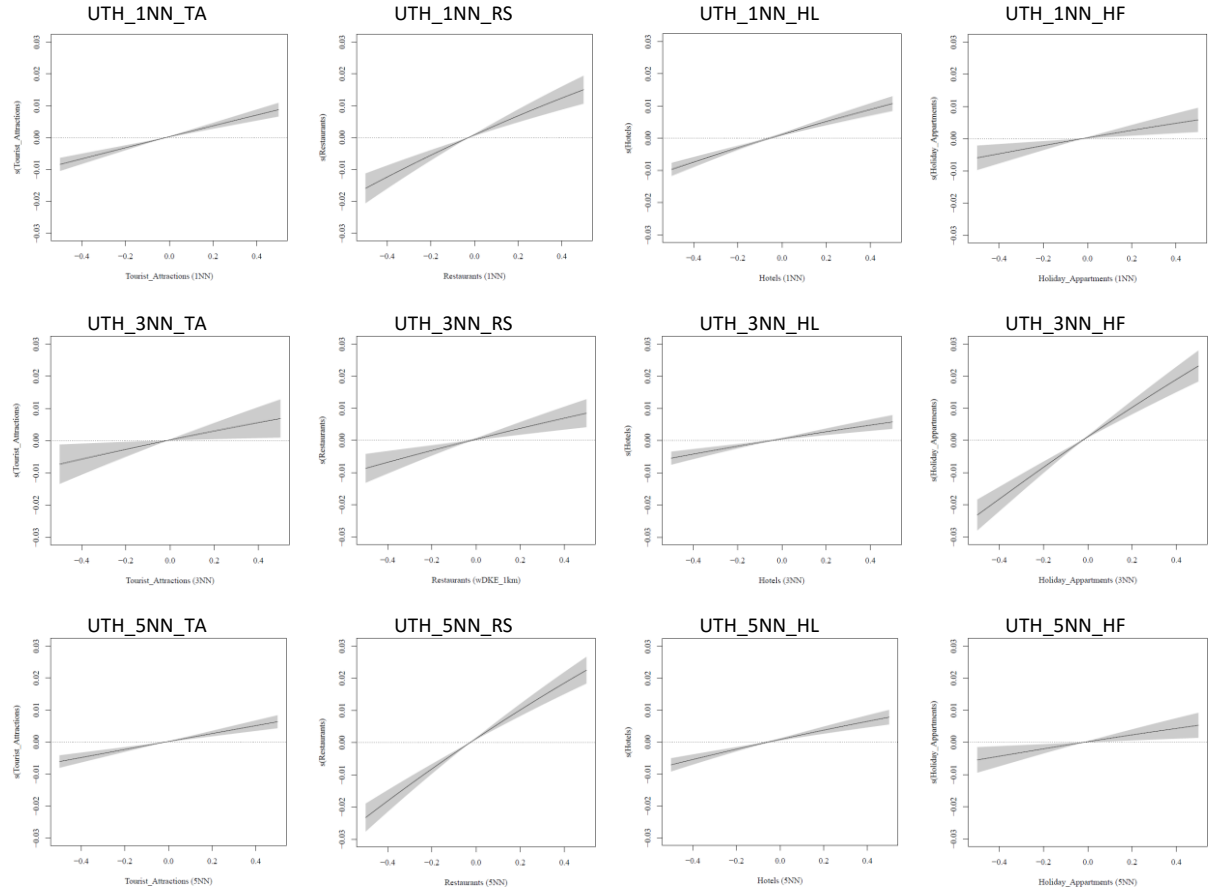
The location effect (*SELA_L)* in the left chart in Figure 3 represents the spatial variability of housing rents, which cannot be explained by any dwelling characteristics, including those spatial characteristics already incorporated. The effect might correspond to the mental map the residents have of their city, for example, concerning the reputation of certain areas or the perceived level of security or livability. This interpretation corresponds to the rather negative values for Marzahn-Hellersdorf (east) and the southern part of Neukölln (south-east). The low values of Spandau (west) might be associated with its distance to the center of Berlin. In contrast, Mitte, Friedrichshain-Kreuzberg and Charlottenburg-Wilmersdorf benefit from the central position. Steglitz-Zehlendorf

[20] *Gründerzei t*refers to the period in Central Europe in the 19th/early 20th Century and was characterized by the architectural style called art nouveau (*Jugendstil*). During this time, several art nouveau villas (*Jugendstilvillas*) were built in Berlin. Several were not destroyed during World War 2 and are generally very attractive places of residence.

(south-west) may be less centrally located, but benefits from its good reputation and a relatively high share of *Jugenstilvillas*.

The right chart in Figure 3 represents the interaction between location and the age (*SELA_L*A*) of a dwelling. High values indicate that an advanced age has a particularly positive effect on rents at this location and vice versa. This can help to distinguish between the effect of different construction styles from the same period of time. The very distinct high values in the eastern part of Marzahn-Hellersdorf represent an area with single- and two-family houses, which were mainly built in the 1980s. In other areas of Berlin, this period is associated with rather low rents, but in this part of Marzahn-Hellersdorf, the interaction effect indicates a positive exception to this relationship. The opposite holds true for the northern part of Marzahn-Hellersdorf, which is dominated by a complex of large-panel system buildings from the mid-1980s. These buildings represent a rather late example of this low-attraction building type and accordingly, the interaction effect shows very low values.

**Figure 4: Effect of urban tourism hotspots (kNN) on housing rents (M7-M9)**

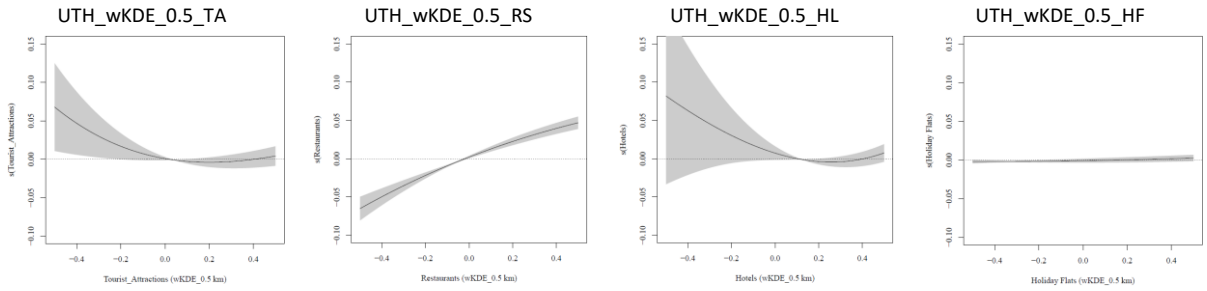



*Effects of UTH on housing rents:* By adding the variables of *UTH* (*tourist attractions, restaurants, hotels and holiday flats*) in the respective *GAM* models, all four models M7-M10 also improve slightly, regarding the explained deviation and minimizing *MSE*, and also show that the single *UTH* affect housing rents. For instance, M9 shows that an increase in the *tourist attraction variable* (*UTH_5NN_TA*) of one standard deviation leads to an increase in housing rents of 1.3 percent, *restaurants* (*UTH_5NN_RT*) have an effect of 4.5 percent, *hotels* (*UTH_5NN_HL*) of 1.6 percent and *holiday flats* (*UTH_5NN_HF*) of 1.0 percent on housing rents. Here again, the results depend on the type of the model and its respective applied approach. However, the results for the *kNN* approach

yield relatively similar results in all three models (M7-M9), particularly by comparing M7 with M9. The splines for all three models and its four categories (Figure 4) show a nearly linear, but partly minimal concave smooth function within the unit of one standard deviation. The steepest functions exhibit the respective splines for *restaurants* (M7 and M9) and the spline for *holiday flats* (M8). Thus, the results also confirm positive effects on housing rents through the four *UTH* categories. This means that a dwelling with a certain number of attractive tourist amenities of the respective category in the proximity, has higher rents than a dwelling without attractive tourist amenities or with unattractive (low score) amenities in its proximity.

The estimates for the *wKDE* approach (M10) show that their effect depends strongly on its respective density value, meaning the effects change with a changing density of *UTH*. In addition, the splines show more distinct non-linear relationships (Figure 5).

**Figure 5: Effect of urban tourism hotspots (wKDE) on housing rents (M10)**

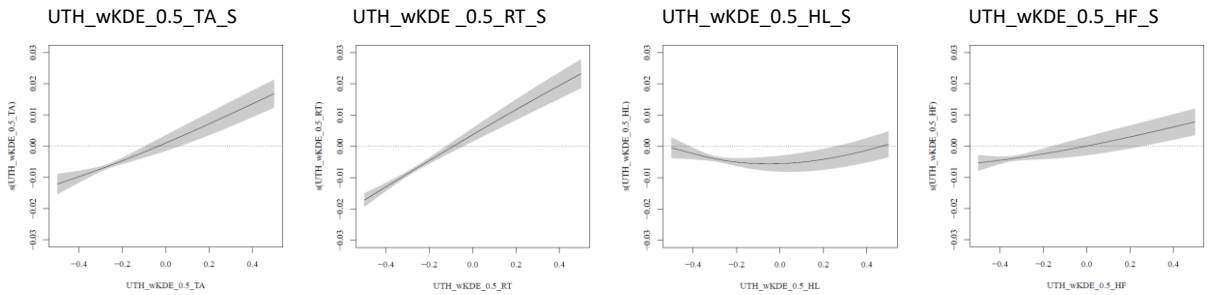



However, as mentioned, there are some multicollinearity problems between single *UTH_wKDE* variables (cf. Appendix 1), so that the results for M10 are presumably biased and thus also to some degree questionable. Nevertheless, when considering the *UTH_wKDE* variables separated from each other in further models, the results of these unbiased models confirm positive effects from the single *UTH* categories on the housing rents (cf. Figure 6). [21]

Here, *tourist attractions (UTH_wKDE_0.5_TA_S)* show with emerging deviations from the mean, a slight concave function with positive slope within the unit of one standard deviation, meaning an effect of *tourist attractions* of about 12.0 percent on surrounding housing rents. *Restaurants* (*UTH_wKDE_0.5_RT_S*) initially depict a slight smaller deviation from the mean and a slight concave function with a positive slope as well, but not as steep as the function of *tourist attractions.* The effect of *restaurants* here is about 7.0 percent on housing rents. The remaining two variables both show a positive slope, but contrary to the first two splines, with slight convex functions. *Hotels* (*UTH_wKDE_0.5_HL_S*) show an effect of about 19.0 percent and *holiday flats* (*UTH_wKDE_0.5_HF_S*) show an effect of about 1.5 percent on housing rents within a bandwidth of 0.5 kilometers. The effects of *UTH* on housing rents increase with an increasing density of amenities.

[21] The same holds for models with separately considered *UTH* variables by using a bandwidth of 1.0km.

**Figure 6: Separated effect of the single urban tourism hotspots (wKDE with bw=0.5km) on housing rents**




Like the OLS results, the results for the GAM models also confirm that the four *UTH* amenities (*tourist attractions*, *restaurants*, *hotels* and *holiday flats*) affect housing rents and that the discussion about increasing rents through tourism in Berlin is reasonable to a certain extent. Most noticeably, the results fuel the debate on the negative effects of *holiday flats* on the Berlin housing market due to rising rents of surrounding rental flats. [22] Even if the exact influence is not definitively quantifiable and the analysis also does not distinguish between the legitimate and illegitimate use of flats (cf. Schäfer and Braun, 2016), it is obvious that *holiday flats* are likely to add value to the surrounding flats.

In total, GAM is more useful in explaining spatial varying parameters determining housing rents than traditional OLS with spatial-invariant and linear relationships. Furthermore, the models show that it is very difficult to accurately measure the real quantity of the respective effects of the *UTH* amenities, and that the results depend strongly on the type of model and its respective applied approach. Intuitively, *tourist attractions* add value to rents, since attractive locations are also appreciated and demanded by locals for instance, because of a good view. *Restaurants* and *hotels*, however, mainly cause shortages of space through increased demand, and *holiday flats* are often located in highly attractive residential areas, where rents are high anyway and if demand the for *holiday flats* increases significantly (as currently through the short-term rental provider) they can additionally stress the housing market. Consequently, there can be direct or indirect effects on the housing market through urban tourism.

### 6.5.3 Model performance and robustness checks

By measuring the goodness of fit with *adj. $R^2$* as a model decision criterion, the respective values clearly show the superiority of the semi-parametric functional form (*GAM*), compared to the conventional parametric functional form (*OLS*). *GAM* outperforms *OLS* regarding the explained deviation by about 13 percentage points on average, and *MSE* decreases by about 25 percent. According to these two performance criterions, the models with *UTH* perform better than without *UTH*, and the models with *UTH_wKDE* perform better than those with *UTH_kNN* (see Table 3).

[22] Again, due to the low number of holiday flats with any rating, the score in this amenity category is composed of the logarithmized number of beds only.

**Table 3: Model performance of the ten models**

| | OLS | | | | |
|---|---|---|---|---|---|
| | **M1 without UTH** | **M2 with UTH 1NN** | **M3 with UTH 3NN** | **M4 with UTH 5NN** | **M5 with UTH wKDE_0.5** |
| N | 68,984 | 68,984 | 68,984 | 68,984 | 68,984 |
| Adj. $R^2$ (%) | 45.8 | 45.9 | 46.1 | 46.2 | 47.7 |
| MSE | 0.040 | 0.040 | 0.040 | 0.040 | 0.039 |
| | **GAM** | | | | |
| | **M6 without UTH** | **M7 with UTH 1NN** | **M8 with UTH 3NN** | **M9 with UTH 5NN** | **M10 with UTH wKDE_0.5** |
| N | 68,984 | 68,984 | 68,984 | 68,984 | 68,984 |
| Adj. $R^2$ (%) | 59.0 | 59.1 | 59.2 | 59.2 | 59.5 |
| MSE | 0.031 | 0.030 | 0.030 | 0.030 | 0.030 |

However, due to multicollinearity problems of the *UTH_wKDE* variables, *kNN* seems to be more appropriate than *wKDE* for this analysis, despite the higher explanatory power of the latter. The *kNN*-variables do not show multicollinearity (see Appendix 1).

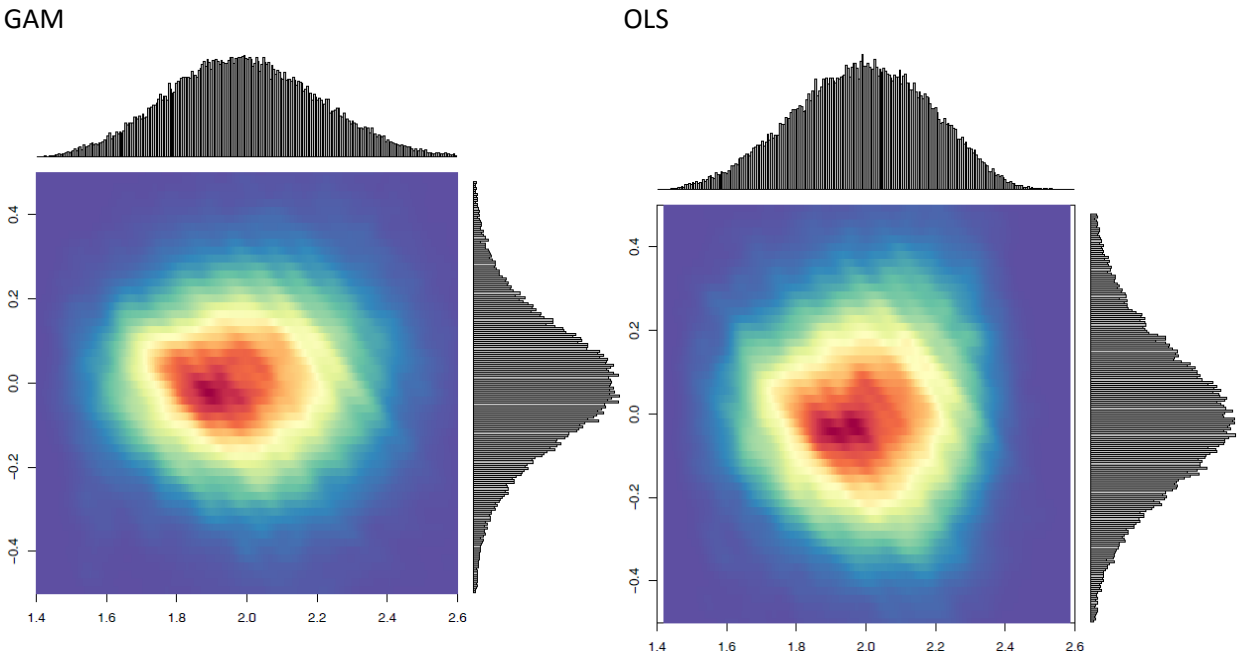
Several further checks of the model assumptions for *OLS*, as well as for *GAM*, were performed and confirm robust models. Regarding the regression residuals, a *Kolmogorov-Smirnov test* does not reject the assumption of normality for both model types. However, a cartographic analysis of the residuals gives a clearer picture of the performance of the two models. Note that a local concentration of positive or negative residuals indicates that certain spatially varying background variables have been neglected in the model. As a sound graphic evaluation, based on the residual of every dwelling, is not possible due to the high number of units, the *Kriging* technique was used to create a continuous surface map of local average residual values (see Appendix 2 and Appendix 3 for M4 and M10). Generally, both methods yield good results in terms of the spatial distribution of residuals, without large clusters of positive or negative values. The two maps further illustrate that residual values are generally lower for the *GAM* in comparison to *OLS*, and particularly that the residual values are distributed in a more compartmentalized way, that is with less spatial autocorrelation.

The analysis of residual values as a function of fitted values is shown in Figure 7. Two-dimensional density values were calculated, because the visual analysis is impeded by the large number of data points. Generally, *OLS* (M4) and *GAM* (M10) produce normally distributed fitted values and residuals, whereby *GAM* performs better in reproducing the whole range of rents including higher values, for which *OLS* fails. Both OLS and *GAM* show some small increase in variability for higher rents.

## 6.6 Conclusion

The analysis confirms that *UTH* (*tourist attractions*, *restaurants*, *hotels* and *holiday flats*) affect the rents of surrounding flats and that the four different types of *UTH* cause different effects, depending on the type of model and its functional form. Consequently, the two established null hypothesis ($H0_1$ and $H0_2$) can be rejected. In addition, the results confirm, due to the construction of the *UTH* variables, which considers the attractiveness of *UTH*, that more attractive *UTH* add more value to the housing rents in close proximity and vice versa. Policy makers and urban planners should be aware of these findings and take them into consideration to avoid further anger regarding affordability and

**Figure 7: Distribution of fitted values and residual values for GAM (M10) and OLS (M4)**




gentrification fears of the local Berlin population. However, it is very difficult to measure the explicit and individual effects (which can be direct or indirect) of the single *UTH* variables and thus to quantify them in a manner that conforms to reality, as the different models with its different approaches show.

The analysis applies the two approaches of *GAM* and *OLS* and shows that the former is a more sound and accurate hedonic tool in explaining spatially varying parameters determining housing rents, than traditional *OLS* with spatial-invariant and linear relationships between regressor and the predictor, as already confirmed in the literature (e.g. Bin, 2004; Cajias, 2014). Moreover, despite lower explanatory power, for this analysis, the *kNN*-approach seems more appropriate than *wKDE* when simultaneously considering the four *UTH* variables, due to multicollinearity of the *wKDE* variables.

It should be mentioned that these results could be different in other cities. Consequently, when the data situation improves, it might be useful to compare different cities with each other. Moreover, it would be worth analyzing, whether this phenomenon holds for different city types (e.g. cultural cities vs. trade fair cities) as well as for cities in other countries, or whether there are differences between the different types of *tourist attractions* (e.g. cathedral vs. museum).

## 6.7 References

Ahlfeldt, G. M., Maennig, W. (2010). Substitutability and Complementarity of Urban Amenities: External Effects of Built Heritage in Berlin, *Real Estate Economics*, 38(2), 285-323.

Bao, H. X. H., Wan, A.T.K. (2004). On the Use of Spline Smoothing in Estimating Hedonic Housing Price Models: Empirical Evidence Using Hong Kong Data, *Real Estate Economics*, 32(3), 487-507.

Barazini, A., Ramirez, J. V., Schaerer, C., Thalmann, P. (2008). *Basics of the Hedonic Price Model*, in: Barazini, A., Ramirez, J. V., Schaerer, C. and Thalmann, P. (Eds.), Hedonic Methods in Housing Markets – Pricing Environmental Amenities and Segregation, pp. 1-12, Springer, New York.

Benson, E. D., Hansen, J. L., Schwartz, A. L., Smeresh, G. T. (1998). Pricing Residential Amenities: The Value of a View, *Journal of Real Estate Finance and Economics*, 16(1), 55-73.

Biagi, B., Faggian, A. (2004). The Effect of Tourism on the House Market: The Case of Sardinia, Ersa Conference – Porto, 25.-29. August 2004.

Biagi, B., Brandano, M. G., Lambiri, D. (2015). Does Tourism Affect House Prices? Evidence from Italy, *Growth and Change*, 46(3), 501-528.

Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions, *Journal of Housing Economics*, 13(1), 68-84.

Bookstein, F. L. (1989). Pricipal warps: Thin-plate splines and the decomposition of deformations, *IEEE transactions on pattern analysis and machine intelligence,* 11(6), 567-585.

Cajias, M. (2014). Spatial effects and non-linearity in hedonic pricing – should we reconsider our assumptions?, Working Paper, 2014.

Conroy, S. J., Milosch, J. L. (2009). An Estimation of the Coastal Premium for Residential Housing Prices in San Diego county, *Journal of Real Estate Finance and Economics*, 42(2), 211-228.

Craven, P., Wahba, G. (1979). Smoothing Noisy Data with Spline Functions, *Numerische Mathematik*, 31(4), 377-403.

Do, A. Q., Grudnitski, G. (1995). Golf Courses and Residential House Prices: An Empirical Examination, *Journal of Real Estate Finance and Economics*, 10(3), 261-270.

Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces, *Revue françaisd'automatique, informatique, recherché opérationelle. Analyse numérique,* 10(83), 5-12.

Eilers, P. H. C, Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties, *Statistical Science*, 11(2), 89-102.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multidimensional probability density, *Theory of Probability and Its Applications,* 14(1), 153-158.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*, 2nd Edition, Marcel Dekker, New York.

Fortheringham, S., Brundson, C., Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley, Chichester.

Füller, H., Michel, B. (2014). Stop Being a Tourist! New Dynamics of Urban Tourism in Berlin-Kreuzberg, *International Journal of Urban and Regional Research*, 38(4), 1304-1318.

Galster, G., Tatian, P., Petit, K. (2004). Supportive Housing and Neighborhood Property Value Externalities, *Land Economics*, 80(1), 33-54.

Geniaux, G., Napoléone, C. (2008). Semi-Parametric Tools for Spatial Hedonic Models: An Introduction to Mixed Geographically Weighted Regression and Geoadditive Models, in: Barazini, A., Ramirez, J. V., Schaerer, C., Thalmann, P. (Eds.). Hedonic Methods in Housing Markets – Pricing Environmental Amenities and Segregation, pp. 101-127, Springer, New York.

Gillard, Q. (1981). The Effects of Environmental Amenities on House Values: The Example of a View Lot, *Professional Geographer*, 33(2), 216-220.

Hamilton, J. M. (2007). Coastal Landscape and the Hedonic Price of Accommodation, *Ecological Economics*, 62(3-4), 594-602.

Hanink, D. M., Cromley, R. G., Ebenstein, A. Y. (2010). Spatial variation in the determinants of house prices and apartment rents in China, *Journal of Real Estate Finance and Economics*, Vol. 45, pp. 347-363.

Hastie, T. J., Tibshirani, R. J. (1990). Generalized Additive Models*, Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Investitionsbank Berlin (Ed., 2013). IBB Wohnungsmarktbericht 2013, online: http://www.ibb.de/portaldata/1/resources/content/download/ibb_service/publikationen/IBB_Wohnungsmarktbericht_2013.pdf (accessed: 04.09.2014).

Kholodilin, K. A., Mense, A. (2012). Internet-Based Hedonic Indices of Rents and Prices for Flats: Example of Berlin, *Discussion Papers of the German Institute for Economic Research DIW*, Berlin.

Knight, J. R. (2008). Hedonic Modelling of the Home Selling Process, in: Barazini, A., Ramirez, J.V., Schaerer, C., Thalmann, P. (Eds.), Hedonic Methods in Housing Markets – Pricing Environmental Amenities and Segregation, pp. 39-54, Springer, New York.

Lu, B., Charlton, M., Fotheringham, A. S. (2011). Geographically Weighted Regression using a Non-Euclidean Distance Metric with a Study on London House Price Data, *Procedia Environmental Sciences*, 7, 92-97

Milon, W., Gressel, J., Mulkey, D. (1984). Hedonic Amenity Valuation and Functional Form Specification, *Land Economics*, 60(4), 378-387.

Pace, R. K. (1998). Appraisal Using Generalized Additive Models, *Journal of Real Estate Research*, 15(1-2), 77-99.

Palmquist, R. (1980). Alternative techniques for developing real estate price indexes, *Review of Economics and Statistics*, 62(3), 442-448.

Papadias, D., Shen, Q., Tao, Y., Mouratidis, K. (2004). Group nearest neighbor queries, *Proceedings of the International Conference on Data Engineering*, 301-312.

Pompe, J. J., Rinehart, J. R. (1995). Beach Quality and the Enhancement of Recreational Property Values, *Journal of Leisure Research*, 27(2), 143-154.

R Core Team (2013). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economy*, 82(1), 34-55.

Rush, R., Bruggink, T. H. (2000). The Value of Ocean Proximity on Barrier Island Houses, *The Appraisal Journal*, 68(2), 142-151.

Schäfer, P., Braun, N. (2016). Misuse through short-term rentals on the Berlin housing market, *International Journal of Housing Markets and Analysis*, 9(3), in press.

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis, *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Sunding, D. L., Swoboda, A. M. (2010). Hedonic Analysis with Locally Weighted Regression: An Application to the Shadow Cost of Housing Regulation in Southern California, *Regional Science and Urban Economics*, 40(6), 550-573.

Tu, J., Xia, Z. G. (2008). Examining spatially varying relationships between land use and water quality using geographically weighted regression I: model design and evaluation, *Science of the total Environment*, 407(1), 358-378.

VisitBerlin (Ed., 2013). Annual Report 2013, online: http://convention.visitberlin.de /en/article/annual-report (accessed: 06.11.2014).

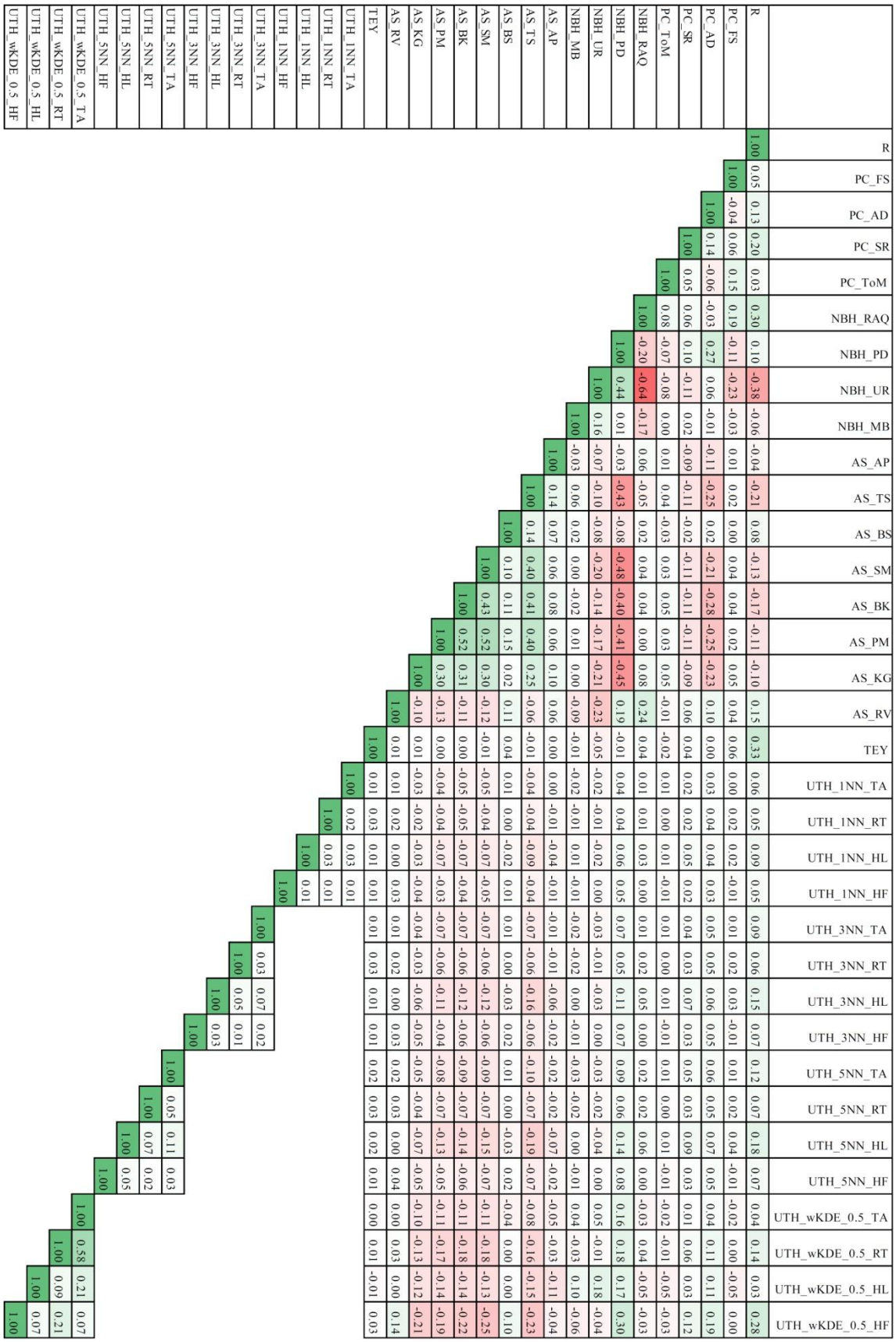
Wood, S. N. (2003). Thin plate regression splines, *Journal of the Royal Statistical Society: Series B*, 65 (1), 95-114.

World Tourism Organization (2012). *Global Report on City Tourism - Cities 2012 Project*, UNWTO, Madrid.

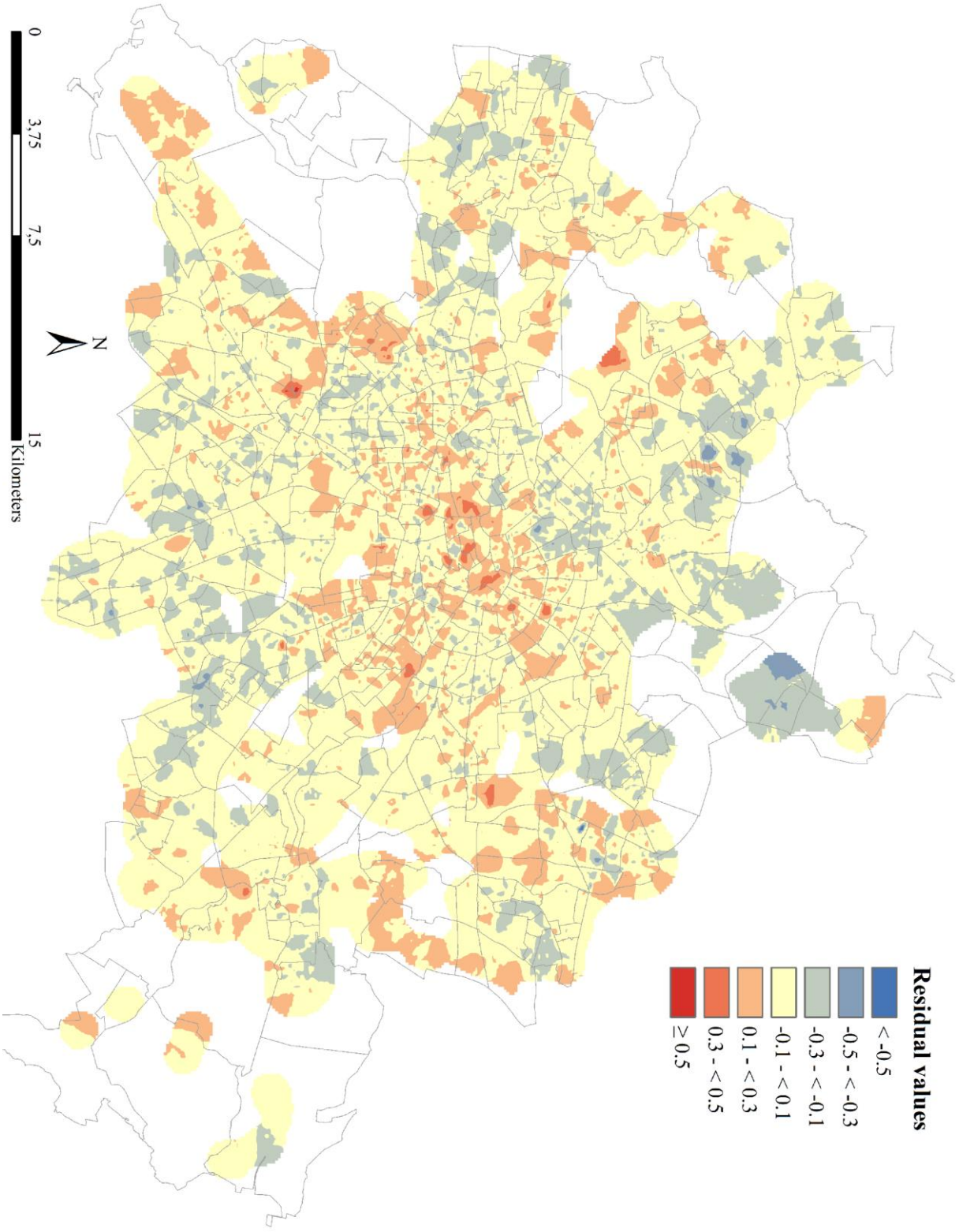
Zheng, S., Kahn, M. E. (2008). Land and Residential Property Markets in a Booming Economy: New Evidence from Beijing, *Journal of Urban Economics*, 63(2), 743-757.

## 6.8 Appendix

**Appendix 1: Correlation-Matrix**

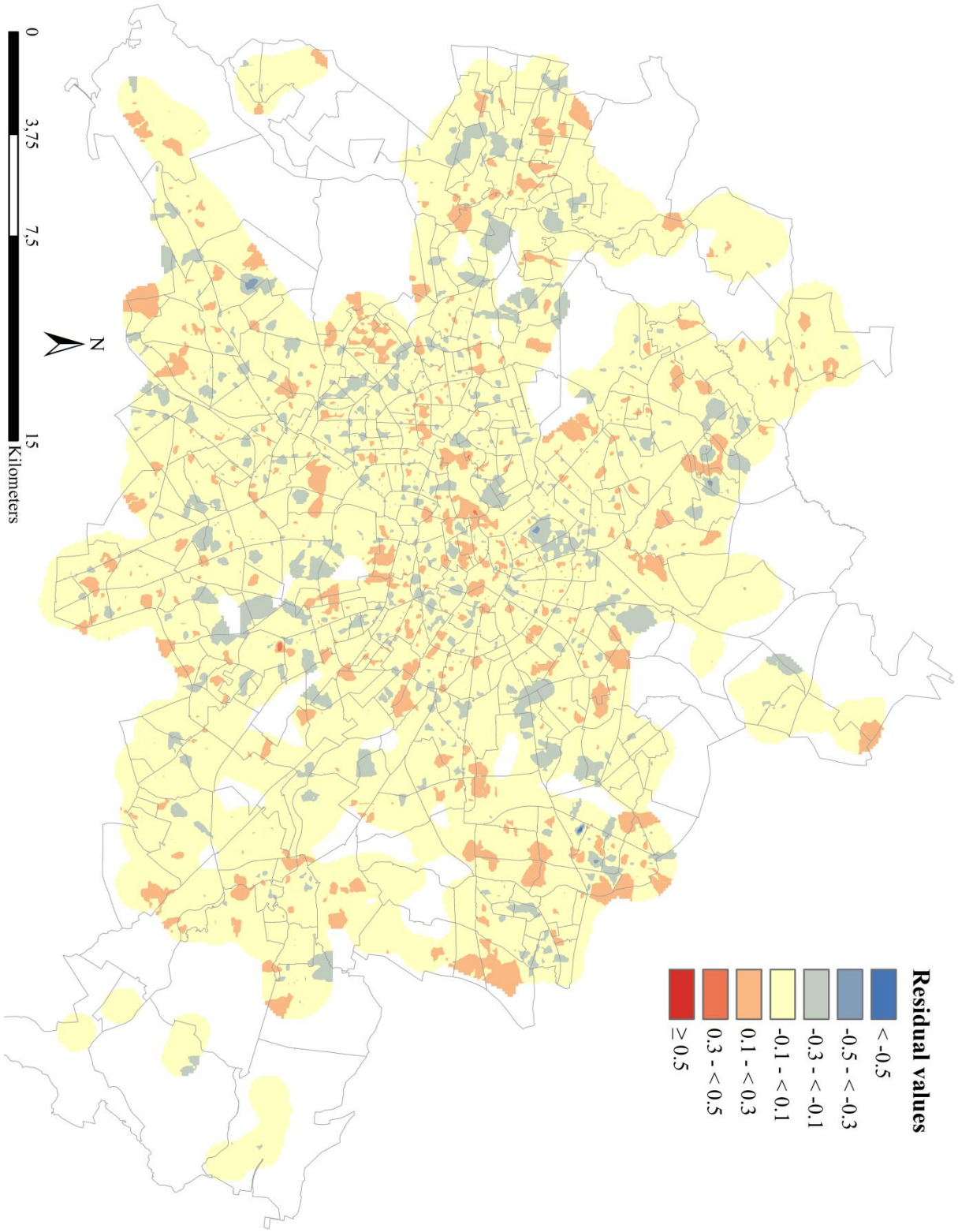


**Appendix 2: OLS Residuals created with Kriging**




Note: Only areas with a certain minimum housing density are shown in the maps of Appendix 2 and Appendix 3.

**Appendix 3: GAM Residuals created with Kriging**



# 7 Conclusion

## 7.1 Executive Summary

**Energy efficiency: Behavioural effects of occupants and the role of refurbishment for European office buildings**

The aim of this paper was to analyze the influence of certain building characteristics and occupant attributes on office buildings. One main focus lay on the effect of refurbishment measures and strengthened legal requirements concerning energy efficiency in recent years.

Our analysis showed that the energy consumption (per square meter) of single-tenant buildings is up to 15% higher, compared to multi-tenant buildings. This result emphasizes the effect of inefficient behavior, when one tenant is responsible for a building as a whole and fails to respond to different conditions and requirements. A modernized *HVAC*-system, which reacts more flexibly, could improve the energy efficiency of single-tenant buildings, as the study found indications of a positive effect of refurbishment dedicated solely to energy efficiency and thermo-physical building characteristics on energy efficiency, even if the effect was not significant due to the small number of observations in this category. In any event, the assumption of a rebound effect connected with solely energy-related refurbishment measures can be rejected. However, refurbishment has a different impact on energy consumption. Major refurbishment measures significantly increase energy consumption, presumably due to additional non-*HVAC* technical and comfort appliances which counteract potential savings. Clearly, a consideration of the interaction effect of building age and refurbishment indicates that the most recent refurbishments place a greater focus on energy efficiency, reflecting stricter building codes and a more efficient building design.

The application of more flexible *GAM* improved the analysis and helped to overcome some restrictions of linear modeling. For example, some extreme outliers regarding Heating Degree Days mediate the results in our linear models, which is not the case for *GAM*. Besides, *GAM* is able to model the saturation effect regarding increasing occupant density. An increased number of occupants per square meter was found to improve energy efficiency, but only up to a certain level. In accordance with the results of other studies (Kahn et al., 2014), *GAM* also rejects the hypothesis of economies of scale. Building size has no significant effect on energy efficiency for small and medium-sized buildings and for very large buildings, energy consumption is even increased, presumably due to less compact building structures.

With investors becoming more and more *CSR*-amenable, it is important to reduce uncertainty about actual energy consumption. The paper demonstrates that a holistic assessment of further sustainability measures besides energy consumption can help to reduce the uncertainty. The assessed intrinsic water consumption in the *GRA* audit was shown to improve the prediction of actual energy consumption in a significant manner.

Finally, it should be noted that the relatively low number of observations in the analyzed sample yields only preliminary results, which could be improved by using a larger dataset of European office buildings. By applying the introduced method to a panel dataset, it may also possible to improve the analysis of the effects of refurbishments.

**Assessment of climatic risks for real estate**

As natural disasters cause billions of Euros of damage to the building stock even today, and climate change will significantly increase these risks, real estate stakeholders need extensive and accurate information about present and future risks, in order to take measures of adaptation. The *ImmoRisk* tool presented in this paper can be regarded as a first step towards offering this information in a quantitative manner. We found that there are no publicly available tools yet that fulfill the requirements of the real estate industry, even if a lot of data on climate and vulnerability can be obtained from a variety of sources. Existing tools tend to neglect relevant hazards or climate change and are constructed rather from the perspective of the insurance than the real estate industry. The proposed trinomial bottom-up approach proposed in this paper is able to connect existing hazard data from regional climate models with property-specific vulnerability functions and data on reconstruction costs, in order to calculate potential monetary losses in an adequate manner. Furthermore, the consideration of future climate conditions enables stakeholders to take well-informed long-term investment decisions regarding location, building types, adaptation measures and portfolio diversification. The *ImmoRisk* tool supports portfolio decisions by enabling the user to create any number of buildings and assess their risk, irrespective of whether these buildings actually exist or are purely hypothetical. Therefore, the tool can be employed by home owners, as well as by institutional investors or communities for urban planning.

The risk assessments conducted with the *ImmoRisk* tool demonstrate the urgent need for comprehensive adaptation measures, as the risk of storms and flooding are expected to increase significantly in the next few decades. Especially locations in northern Germany face a particularly strong increase in storm risk and should be prioritized when institutional investors consider adapting parts of their portfolio. The next research step, which is already being conducted by the authors, is the nationwide extension of the locations covered beyond the 15 pilot locations of the initial *ImmoRisk* tool. The future availability of necessary data rises and falls with an improved coordination of research on hazard data and vulnerabilities, which should be based on the same relevant hazard variables. The *ImmoRisk* project is an initial step towards more awareness and a mutual understanding of the common future research agenda. It was, not least due to this success, evaluated very positively by the German federal principal and awarded with the *Immobilienforschungspreis 2013* by the *GIF, Germany Association for real estate research*.

**The analysis of customer density, tenant placement and coupling inside a shopping centre with GIS**

The aim of this paper was to demonstrate to what extent the theoretical and practical research on shopping centers can be improved by applying *GIS* and, in some fields of interest, its combination with on-site customer surveys. It was shown that *GIS* offer a variety of methods that can be applied by both researchers and center managers.

From the perspective of theoretical shopping center research, *GIS* enables an improved analysis of many proposed hypotheses about customer behavior and spatial structures of tenant placement. The behavioral aspect entails, for example, the routes that customers choose during their shopping trip. *GIS* offers, in combination with the conducted on-site customer survey, the perfect tool for analyzing these movements and, unlike purely camera-based methods of capturing customers, can also consider socio-demographic characteristics and detailed information on behavior at the point of

sale. The proposed decline in customer density with the distance from the center of the mall (Cartel and Vandell, 2005; Carter and Haloupek, 2000) can be tested regarding different definitions of this center, and the analysis can be extended by applying regression models that consider other spatial variables like the distribution of shop size or retail categories. In turn, the question of dispersed or concentrated retail categories requires valid measurement methods. The paper introduces the *Variable Clumping Method* to shopping center research and offers a probability-based measure of category concentration that enables more sophisticated conclusions to be drawn than existing measures like the *Herfindahl index*. Furthermore, *GIS* can be used to approach the complex issue of coupling and enables theoretical and graphic analysis that can also consider a wide range of socio-demographic differentiation, due to the information obtained from the conducted customer survey. In this context, the paper introduces the concept of *relative* coupling, which offers more sophisticated insights into the coupling conditions of individual shops than the traditional concept of *absolute* coupling.

Also, the center management can benefit from an application of *GIS* and the methods introduced in this paper. The analysis of customer flows can help to identify problematic areas of the center with low customer density and resulting low revenues. The socio-demographic differentiation of customer flows might help to improve the tenant mix or tenant placement regarding certain customer groups, by attracting them to certain parts of the center. Furthermore, specific marketing activities can be launched to improve the situation. The detailed analysis of coupling behavior can be regarded as an additional tool for the center management to control customer flows. For example, a more dispersed distribution of strongly coupling shops might help to lead customer flows into areas of low density and to generate more shared and suscipient business (see Nelson, 1958).

**Variable Clumping Method and Mean-k-Nearest-Neighbor Method - Introducing two new approaches to retail concentration measurement to shopping center research**

The question, of whether shops from the same retail category should be located in a dispersed or a clustered manner, has been widely discussed in the shopping center literature. The purpose of this paper was not to answer this question, particularly as Yuo and Lizieri (2013) demonstrated that the correct answer seems to depend on the structural complexity of a shopping center and the consequent expenses of customers trying to access a certain number of shops from the same category. Despite the large number of studies on category concentration, the actual measures for its assessment lack theoretical foundation or have practical shortcomings. For example, the widely applied *Herfindahl index* only assesses the total number or sales areas of a certain category in relation to the respective total numbers, without even looking at their actual spatial distribution. Other approaches are based on the analysis of mean distances of shops from the same and from different retail categories, without providing any further theoretical insights into these measures. In order to improve this situation of an insufficient methodology, we introduced the *Variable Clumping Method* (*VCM*) (Okabe and Funamoto, 2000) into shopping center research and describe in considerable detail how the method can be applied to the specific situation of retail concentration within a shopping center. Accordingly, this paper goes beyond the more results-oriented presentation of *VCM* in our other paper.

*VCM* offers a probability-based approach for the analysis of retail concentration on a multi-scale level and enables the explicit pinpointing of individual clusters of different size. By comparing the

empirically determined number of clumps for several spatial scales with the corresponding expected numbers for a random distribution of shops, *VCM* enables propositions to be made about the statistical significance of single clusters and yields a much more detailed description of actual conditions. Hence, its application could generate more reliable data on retail concentration that can be applied in further research with regression studies on factors explaining shopping center rents. The paper presents a detailed methodology for the implementation and interpretation of *VCM* which can be applied to other shopping centers, as well as to other fields of research such as the distribution of downtown retail agglomerations.

The paper further introduces the *mean-k-nearest-neighbor method*, which provides a less computation-intensive approach than *VCM* for the analysis of concentration tendencies of retail categories. Its application to some generic distributions of shops demonstrate its principal methodology and validity, but nonetheless, further research will be necessary to study its statistical characteristics and behavior in the case of more complex floor plans.

**Do urban tourism hotspots affect housing rents - Evidence from Berlin**

This paper focuses on the Berlin rental market and applies different regression models to identify the main factors driving housing rents, including building characteristics and location. Location variables comprise the distance to certain urban infrastructures and rivers, the situation of a building in a certain area of the city and especially the position relative to four categories of more or less tourist-related amenities (tourist attractions, hotels, holiday flats and restaurants). The paper is motivated by a lack of sophisticated studies focusing primarily on Berlin and by the increasing importance of urban tourism there, with a growth of overnight stays of about 10% per year since 2006. The corresponding effects on rents are subject of heated debate, as the tourist infrastructure could expand at the expense of space for housing and thus trigger gentrification effects.

The analysis is based on several databases which were compiled specifically to improve the analysis. Building characteristics and rental data were received from Germany's leading real estate portal *Immobilienscout24*, socio-demographic data was obtained from municipal statistics, information on urban infrastructure from the open data platform *OpenStreetMap* and data on location and the rating of tourist amenities from the travel website *TripAdvisor*.

One main finding of the paper is not primarily related to the effects of tourism amenities, but to questioning what causes the observed strong rent increases in recent years. The work shows that only a small share of increased rents can be explained by improved building characteristics and location effects. The major part seems to be related to increased demand or other market factors. Interestingly, there seems to be growing investor interest in Berlin's high-price districts. The number of apartment rented in the analyzed period increased approximately proportionally with the regression coefficient of the respective district. By controlling this interaction effect between district and time, we found a significant increase of rents that is not explained by quality improvements in the districts Mitte, Frierichshain-Kreuberg and Neukölln. The absence of significant interaction effects in the other districts suggests that the observed rent increases there are indeed primarily a consequence of improved building qualities.

From the viewpoint of methodology, the application of *GAM* yielded distinctly greater explanatory power than linear models. For instance, the size of living area and the building age display non-linear

effects. Unsurprisingly, the lowest rents, all else equal, where observed for building ages of about 40 years, which is dominated by large-panel system buildings. The consideration of location and its interaction effect with building age as distinct variables in *GAM* (‘*geoadditive models’*) enabled the estimation of disregarded spatial effects like the reputation of certain areas, or the effects of different construction styles from the same building period.

The analysis of effects of tourist amenities first of all required some preprocessing of data from *TripAdvisor* to generate the necessary variables for the regression models. Each individual amenity was assigned a score based on the number and quality of ratings on *TripAdvisor*. In combination with the location of the amenities, which was also collected from *TripAdvisor*, a geodatabase was built. We tested two different approaches for creating the regression variables, based on *kernel density estimators* and a *nearest neighbor approach*. Each model was further diversified regarding kernel density bandwith, or the number of nearest neighbors included for the creation of the regression variables. As the kernel density neighbors were subject to collinearity problems, we propose using the nearest neighbor approach for future analyses on this topic. The studied tourist amenities show a clear positive correlation with rents, but it cannot be ruled out that it is high-pricing areas that attract these amenities instead of the amenities influencing the rents. In any event, it could be shown that the amenities provide some, albeit small, additional explanatory power to the models.

## 7.2 Final Remarks

Coping with climate change and ensuring sustainable economic conditions are two of the main challenges of our time. Mitigating greenhouse gas emissions, adapting building structures to more severe natural hazards, taking legal measures against a scarcity of affordable housing, as well as managing shopping centers against the background of changing customer behavior and new competitors - all these targets are associated with economic considerations and require profound information about and knowledge of the involved structures and processes. The aim of this dissertation was to shed light on these issues from a scientific perspective. As a geographer, my main focus concerning this objective was the exploitation of spatial data in combination with sophisticated statistic methods including GIS. The above papers describe how spatial information and geostatistic methods can be applied to very different spatial scales and questions - from the interior of shopping centers and office buildings to housing rents at the urban level, through to the assessment of natural risks in Germany. I hope that not only researchers, but also practitioners in the field of real estate will be able to draw some inspiring conclusions from the introduced methods and results. This will hopefully help then to formulate their business activities and planning decisions more sustainably, which includes ecological, economic and social sustainability.

The ecological dimension is represented primarily by the paper on energy efficiency in office buildings. A better understanding of technical and behavioral drivers of energy consumption is fundamental, if the real estate industry is to make a further step towards mitigating climate change.

Climate change might be an ecological topic, but its consequences in terms of more frequent or intense natural hazards concern primarily our economic system, and the need for well-informed adaptation measures. All stakeholders, private household, engineers, investors, communities and the insurance industry, can benefit from improved information regarding present and future natural risks. The research activities within the different fields should be better coordinated and results should be processed and communicated in a more user-oriented manner. The *ImmoRisk* tool can be

regarded as a first step in this direction and it is notable that its general methodology can be transferred to other countries that might be even more affected by climate change than Germany.
The question of how shopping center managers can improve structures and processes within their facilities also affects the economic dimension of sustainability regarding the aspiration to make business concepts of shopping centers future-proof against the backdrop of changing customer attitudes and new competitors.
Finally, the topic of increased rents is not only an economic one, but in my opinion, in fact concerns the central idea of what is meant by social sustainability. A lack of affordable housing is associated with social tensions and endangers the peaceful coexistence of individuals. If the development of rents becomes the plaything of pure speculation in some Berlin districts, it is questionable whether market dynamics still act in the best interests of citizens. A scientific understanding of the underlying forces that influence rents can hopefully contribute to dealing with these issues and guiding politics towards the appropriate measures.

## 7.3 References


Carter, C. C., Haloupek, W. J. (2000). Spatial Autocorrelation in a Retail Context, *International Real Estate Review*, 3(1), 34-48.

Carter, C. C., Vandell, K. D. (2005). Store Location in Shopping Centers: Theory and Estimates, *Journal of Real Estate Research*, 27(3), 237-65.

Kahn, M. E., Kok, N., Quigley, J. M. (2014). Carbon Emissions from the Commercial Building Sector: The Role of Climate, Quality, and Incentives, *Journal of Public Economics,* 113, 1-12.

Okabe, A., Funamoto, S. (2000). An exploratory method for detecting multi-level clumps in the distribution of points: a computational tool, VCM (variable clumping method), *Journal of Geographical Systems*, 2(2), 111-120.

Yuo, T. S.-T., Lizieri, C. (2013). Tenant Placement Strategies within Multi-Level Large-Scale Shopping Centers, *Journal of Real Estate Research*, 35(1), 25-52.