Article

Urban Analytics and City Science

EPB: Urban Analytics and City Science
2020, Vol. 47(3) 400–416
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2399808318777500
journals.sagepub.com/home/epb
SAGE

Combining tacit knowledge elicitation with the SilverKnETs tool and random forests – The example of residential housing choices in Leipzig

Sebastian Scheuer
Humboldt-Universität zu Berlin, Geography Department, Germany

Dagmar Haase
Humboldt-Universität zu Berlin, Geography Department, Germany;
Helmholtz Centre for Environmental Research–UFZ, Department of Computational Landscape Ecology, Germany

Annegret Haase
Helmholtz Centre for Environmental Research–UFZ, Department of Urban and Environmental Sociology, Germany

Nadja Kabisch
Humboldt-Universität zu Berlin, Geography Department, Germany;
Helmholtz Centre for Environmental Research–UFZ, Department of Urban and Environmental Sociology, Germany

Manuel Wolff
Humboldt-Universität zu Berlin, Geography Department, Germany; Helmholtz Centre for Environmental Research–UFZ, Department of Urban and Environmental Sociology, Germany

Nina Schwarz
Helmholtz Centre for Environmental Research–UFZ, Department of Computational Landscape Ecology, Germany

Katrin Großmann
University of Applied Sciences Erfurt, Department of Urban and Spatial Planning. Germany

Abstract

Residential choice behaviour is a complex process underpinned by both housing market restrictions and individual preferences, which are partly conscious and partly tacit knowledge. Due to several

Corresponding author:
Sebastian Scheuer, Geography Department, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.
Email: sebastian.scheuer@geo.hu-berlin.de

limitations, common survey methods cannot sufficiently tap into such tacit knowledge. Thus, this paper introduces an advanced knowledge elicitation process called SilverKnETs and combines it with data mining using random forests to elicit and operationalize this type of knowledge. For the application case of the city of Leipzig, Germany, our findings indicate that rent, location and type of housing form the three predictors strongly influencing the decision making in residential choices. Other explanatory variables appear to have a much lower influence. Random forests have proven to be a promising tool for the prediction of residential choices, although the design and scope of the study govern the explanatory power of these models.

Keywords
Residential choice, random forest, tacit knowledge, knowledge elicitation, data mining

Introduction

Managing complex human–environmental interactions requires understanding and knowledge about the dynamics of these systems. To do so, increasingly huge amounts of data are used, creating numerous opportunities and challenges for analysis (Einav and Levin, 2014). Data mining is becoming a focus point of research to cope with such large data sets. Data mining can be understood as: (i) a process of data-driven identification of novel patterns in data, i.e., knowledge discovery; and (ii) the elicitation of predictive models, i.e., knowledge formalization (Rokach and Maimon, 2015; Witten et al., 2011).

The use of big data has considerable potential for studying human-environmental interactions. However, analysing human behaviour with big data – according to Bharwani et al. (2015) a challenging task in complex, high- and multi-dimensional settings – has the drawback that a multitude of factors could potentially influence an actor's decision, whilst information on the drivers underlying this decision are potentially not included in the data. Thus, it is hard to identify the reasoning behind a decision, e.g., in terms of individual preferences, choices and their underpinnings. These underpinnings are commonly referred to as tacit knowledge, which is unconscious, abstract, tied to personal experiences and often unvoiced (Raymond et al., 2010). Being implicit to individuals, tacit knowledge requires formalization to be put into operation.

Experimental settings have been used in the past in various fields such as psychology and economics to control for attributes and confounding factors. Methods such as choice experiments or conjoint analysis have also gained popularity in the study of human–environmental systems (Bartkowski et al., 2015). They do not explicitly ask for reasons underlying a decision, but rather infer them from choices made in hypothetical situations. Choice experiments thus derive stated preferences (SP) by asking participants to choose their preferred option from a given set of alternatives that are described with certain attributes (Gustafsson et al., 2001). This is in stark contrast to more common survey methods such as (semi-)structured interviews, surveys or online voting tools, which ask respondents for their motivation or preferences directly, e.g., based on Likert scale evaluation, thus requiring participants to verbally describe or quantify their preferences, implying that respondents are acutely aware of their motivations and attitudes that guide their decisions. In the case of unarticulated tacit knowledge, these requirements are typically not met (Raymond et al., 2010).

SP approaches are also widely used for modelling residential choice (Timmermans et al., 1992). McFadden (1978) described the residential choice issue as one of maximizing utility, where each alternative is described by attributes such as accessibility to work, shopping, and schools, neighbourhood quality, availability of public services, costs, number of rooms or provided appliances and where individuals' perceptions of dwelling units impact their decision. Kim et al. (2003) highlight the importance of accessibility and transport-related attributes for residential location choice. Bhat and Guo (2004) assessed the impact of residential zones' characteristics (size, population, density, etc.). Walker and Li (2007) studied the role of lifestyles in choice behaviour by using latent class choice modelling, and Tu et al. (2016) estimated the impact of urban green space availability. Stokenberga (2019) quantified the importance of informal support networks, whereas Ibraimovic and Masiero (2012) studied the effects of ethnic neighbourhood composition and ethnic preferences. Yu et al. (2017) analysed residential mobility as intertemporal location choices, i.e., changes in location over time due to dependencies of manifold time-invariant and time-variant factors.

Clearly, residential choice interacts in a complex way with various (spatial) determinants. It can be considered as a decision with long-term consequences that draws on a complex set of criteria including stated and tacit preferences and constraints. Furthermore, residential choice affects, directly and indirectly, market developments, the built environment, networks, urban form and land-use (Sener et al., 2011). Consequently, for stakeholders, residential choice modelling is an important tool for urban, transport and land-use planning (Kim et al., 2003), first and foremost for steering actual and future housing demand and supply (Stokenberga, 2019). However, the complexity of interactions in residential choice is difficult to conceptualize, especially as preferences may undergo rapid alterations.

Combining SP choice experiments with methods of statistical and/or machine learning allows patterns of tacit knowledge to be explored with the help of rigorous statistical techniques. Combining these two approaches is the core element of a computer-aided knowledge elicitation approach (KnETs) described by Bharwani (2006). Conceptually, the KnETs process presents a participatory knowledge elicitation approach, complementary to more conventional conjoint analysis that relies on choice experiments to identify the criteria motivating human decisions, and to uncover unvoiced prioritizations or unconscious evaluations (Bharwani, 2006).

The KnETs process includes the four principal stages following Wood and Ford (1993) and Wooten and Rowley (1995) as described in Figure S1 that combine qualitative and quantitative methods and result in a formal knowledge representation (Bharwani, 2006; Bharwani et al., 2015). KnETs uses a controlled, structured, interactive and iterative interviewing method that results in a 'game' being played, so that factors that can potentially influence a decision are constrained, and large datasets be gathered. The KnETs approach has been used in a variety of case studies including crop choice modelling in South African communities (Bharwani et al., 2005) and Cameroon (Bharwani et al., 2015) and the EU FP6 NeWater project on adaptive river basin management (Kemp-Benedict et al., 2010).

The study presented in this paper advances the KnETs approach in two innovative ways. From a technical perspective, it seeks to improve KnETs by proposing SilverKnETs as a new software tool to conduct computer-aided, iterative interviews building upon KnETs experiences. Methodically, KnETs is sought to be improved by using random forests (RFs) as an alternative machine-learning approach for the statistical analysis of interview outcomes. For showcasing, a case study to investigate residential choice behaviour is presented. As described above, residential housing choice, particularly in the urban space, is a prime example of dynamic and complex decision-making underpinned by personal preferences intertwined with social, economic, environmental and spatial aspects and interactions. Consequently, to enable planners to make informed decisions, the relevant preferences, motivations and criteria underpinning residential choices of market participants need to be identified and formalized. It is in this regard where data mining and particularly RFs are considered as a major methodological opportunity.

Advancing KnETs in form of SilverKnETs

Looking at Figure S1, the KnETs process is strongly dependent on software tool support in two contexts, i.e., the interview context and the data mining context. Originally, a JAVA-based survey tool developed by Michael Fischer, University of Kent, UK, provided this tool support (sourceforge.net/projects/knets/). Several limitations can be identified for this original version of KnETs including: (i) missing capabilities for the generation of random values; (ii) lack of internal means of scenario validation; (iii) lack of data models to distinguish between views – i.e., what is shown to the participant on screen, also in light of localization aspects – and export, i.e., how data are encoded internally for the facilitation of data mining; (iv) lack of conjoint interviews; (v) lack of web-based capabilities to conduct interviews remotely and in an unsupervised manner. Commercial products exist that overcome these limitations, however, at potentially inhibiting licensing costs, thus giving rise to a novel adaption of KnETs.

SilverKnETs is developed to be freely available and to address most of KnETs' limitations. It features conjoint interviews and an internal scenario validation engine, which allows setting certain restrictions on the combinations of predictor values during scenario generation to include presumptions, limitations or hypotheses on the problem domain at hand. SilverKnETs also separates the view model from the data model, so that the way information is presented on the computer screen is independent from the way values are recorded for the subsequent application of data mining techniques, thereby enabling user-centric, localized surveys. SilverKnETs follows a loose-coupling approach, thus enabling the export of data into a standardized file format so that knowledge discovery and formalization can be conducted using any statistical software.

Application of SilverKnETs to elicit residential housing choice behaviour in the city of Leipzig

In this case study, SilverKnETs is used to elicit (tacit) knowledge-driving residential choices within the city of Leipzig, Germany. Leipzig represents a highly dynamic housing market with continuously high residential mobility (Welz et al., 2014; Wolff et al., 2016). A phase of considerable shrinkage from the 1960s until about the end of the 1990s was followed by a period of stabilization, succeeded by the onset of dynamic growth of 2% p.a. from 2010 onwards (Haase and Rink, 2015; Wolff and Haase, 2015). These rapidly shifting population dynamics pose difficulties for planners to estimate future residential demand in the city. Various questionnaire surveys and qualitative interviews were used repeatedly to gain knowledge on housing choices and migration behaviour in Leipzig (Grossmann et al., 2015; Haase et al., 2012a; Stadt Leipzig et al., 2016). However, as outlined above, these common survey methods may not sufficiently elicit underlying (tacit) preferences. Analysing the outcomes of this decision-making process in the form of patterns and changes in the housing market or net-migration flows using census data and municipal statistics may also omit factors of relevance. Consequently, the advanced KnETs process in the form of the SilverKnETs tool is reasonable for application to this case.

Preparatory domain exploration

Following Ettema (2010), factors that influence residential choice are relatively stable and include housing attributes, neighbourhood attributes, as well as accessibility and transport-related characteristics. In this case study, the variables *rent*, *total area* of the flat as well as the *number of rooms*, the *location* within the city, *house type* of the apartment building, presence of *neighbourhood amenities* as well as *furnishing* of the apartment were elicited accordingly during several expert rounds as part of domain exploration (Table 1). The choice of these attributes was

Table 1. Housing, neighbourhood and household attributes (predictors) with corresponding domain values.

Type	Predictor	Description	Domain values
Housing and neighbourhood attributes	Rent	Categorized cost of rent including heating (EUR)	(0;300][a], (300;400], (400;500], (500;700], (700;900], (900;1,200], (1,200;1,500], (1,500,∞)[a]
	Location	City district	Centre, East, South, West, Gohlis, Grünau, Outskirts, Hinterland
	Rooms	Number of rooms	[1], [2], [3], [4], [5;∞)
	Total area	Categorized apartment size (total, m^2)	(0;20], (20;40], (40;60], (60;80], (80;100], (100;∞)
	House type	Structure type of the apartment building	Wilhelminian (NR), Wilhelminian (PR), Wilhelminian (FR), GDR prefabricated block (NR), GDR prefabricated block (PR), GDR prefabricated block (FR), Post-reunification building, Detached house (NR), Detached house (PR), Detached house (FR)
	Neighbourhood amenities	Presence of neighbourhood amenities in the vicinity of the apartment	Bar/restaurant, family-friendly neighbourhood, main road, park (urban green), pharmacy, school, shopping, multi-cultural neighbourhood
	Furnishing	Furnishing features of the apartment	Bathtub, individual bathroom, courtyard and/or garden, fitted kitchen, lift, parquet floor, modern insulation
Household attributes	Income	Categorized net income (EUR)	(0;600], (600;1100], (1100;1600], (1600;2100], (2100;2600], (2600;3100], (3100;∞)
	Employment status	Employment status	Employed full time, Employed part-time, In education, unemployed, not employed due to other reasons, pensioner, not applicable
	Qualification	Highest level of education	University degree, University of applied sciences degree, Master (craftsman) degree, skilled worker, Teilfacharbeiterabschluss, In education, without professional qualification, not applicable
	Age	Categorized age	[18–20), [20–30), [30–40), [40–50), [50–60), [60–70), [70–80)

GDR: German Democratic Republic; NR: not renovated; PR: partially renovated; FR: fully renovated.

[a] Denotes class limits.

further based on long-term experiences of surveying seeking to understand residential decisions and associated housing demand, and the population dynamics in cities including Leipzig (Grossmann et al., 2015; Haase et al., 2012a, 2012b; Haase and Rink, 2015; Welz et al., 2017). An indicator set for reurbanization processes (Kabisch et al., 2010) and expert experiences from modelling additionally framed this selection (Haase et al., 2010; Lauf et al., 2012).

As shown in previous studies, the aforementioned attributes tend to interact with household characteristics, i.e., socio-demographic factors (Ettema, 2010). Consequently, predictor importance and individual preferences may vary considerably, e.g., between different groups of income or age (Angelini and Laferrère, 2012; Kim et al., 2003; Park and Kim, 2016), due to internal household dynamics (López-Ospina et al., 2016), or over time based on experiences, life plans, life events and (external) shocks (Bajari et al., 2013; Clark and Huang, 2003; Yu et al., 2017). To reflect on this heterogeneity in preferences, the household attributes *income*, *employment status*, *qualification* and *age* were also included in the case study (Table 1).

Data sampling using interactive interviews

The data sampling was carried out in the form of interactive interviews within a two-week period in March 2015 at various inner-city locations in Leipzig including the city centre, the university campus, residential areas of different type and large shopping malls. Additionally, a sample consisting of residents who recently changed place of residence was collected in the various local city offices of Leipzig where inhabitants register. The selection of sampling locations is based on former interview experiences with the aim of receiving a certain rate of response to allow for further analysis (Welz et al., 2014, 2017). In the interviews, scenarios were generated iteratively, and respondents had to accept or decline each alternative (Figure S2). Each scenario represents a potential apartment similar to an advert, created by randomly drawing factors from the set of domain-specific values for each predictor (Table 1), and controlled for by SilverKnETs to eliminate non-representative options.

Knowledge formalization and evaluation

Predictive data mining models for classification and/or prediction tasks include neural networks, Bayesian networks, support vector machines, single classification and regression trees (CART) as well as RFs, i.e., ensembles of CART (Breiman, 2001; Lausch et al., 2015; Wright et al., 2016). These methods are typically ascribed to supervised learning methods, i.e., they rely on a pre-specified target attribute that should be predicted by a set of independent predictors (Hastie et al., 2009; Rokach and Maimon, 2015). RF classifiers are widely used in research, such as in remote sensing, and were found to outperform CART and common regression methods (Antipov and Pokryshevskaya, 2012; Belgiu and Drăguţ, 2016; Rodriguez-Galiano et al., 2012). Thus, RFs promise to be a competitive and efficient machine-learning approach. Other advantages of RF are commonly seen as: (i) their model-agnostic nature; (ii) handling of large numbers of mixed – qualitative (categorical) and quantitative – predictors; (iii) their robustness to outliers and (iv) their capability of effectively dealing with very large data sets. Further benefits include their internal error estimate (out-of-bag error, OOB), and their internal estimate of variable importance (Breiman, 2001; Rodriguez-Galiano et al., 2012).

Due to this perceived superiority, this case study uses RF to evaluate their performance in residential choice prediction. In this regard, the presented case study seeks to improve KnETs methodically, with the latter relying on CART (Bharwani et al., 2015). To the knowledge of the authors, RFs have rarely been used in the residential choice context. Instead, regression methods – e.g., multinomial logit regression and nested logit regression models – are more commonly used (Yates and Mackay, 2006). Antipov and Pokryshevskaya (2012) have used RF to determine the importance of mixed predictors on housing prices in St. Petersburg, Russia, where RF outperformed other methods such as CART, neural networks, or multiple regression analysis.

Results

RF models

In the following, the performance of RF models that include only housing and neighbourhood attributes is compared with models that additionally account for heterogeneity in residential choice by including the household attributes listed in Table 1. To build the former models, a total of 7712 scenarios were used that have been sampled from 199 individual respondents; the median number of scenarios per respondent is 30. To build the latter models, only 7450 scenarios were used, excluding 262 cases from 21 unique respondents due to entirely missing household attributes.

RF generation was carried out in the R statistics software using packages 'randomForest' (Liaw and Wiener, 2002) and 'randomForestSRC' (Ishwaran et al., 2008). For RF training, cases were randomly split into a training data set (80%) and a test data set (20%). The majority class, i.e., no cases, indicating the rejection of a scenario, outnumbers the minority class, i.e., the class of interest, in a ratio of approximately 1:12, with 7144 no cases to 586 yes cases. To deal with this imbalanced data, as suggested by Chen et al. (2004), a downsampling approach has been used on the majority class in the training set. All RFs were grown using 600 trees, a number deemed sufficient to let the OOB error converge towards the estimated true prediction error.

Variable importance

For identifying important predictors, RF provide built-in means of estimating variable importance (Breiman, 2001). In the 'randomForest' package, this measure is the mean decrease of accuracy (MDA). It can be assumed that the more important the variable is, the greater the MDA (Hastie et al., 2009). The 'randomForestSRC' package implements an alternative importance measure based on the minimal depth of a predictor's maximal subtree (Ishwaran et al., 2011). Maximal subtree refers to any given subtree of a CART for which no higher split on the predictor is present. The 'distance' of this maximal subtree to the root node where distance = 0 is called the minimal depth. Hence, early split criteria are indicated by smaller minimal depths, and the smaller the minimal depth, the more important is a predictor's role in the decision-making process (Wright et al., 2016). Using either importance measure, *rent*, *location* and *house type* are the three most-important predictors (Table 2). If included, household attributes appear to be of only moderate importance, with *income* and *qualification* being the most important household attributes.

Predictor interactions

In the following, using randomForestSRC, the probability for the prediction of a specific outcome (class) is investigated more closely for the three most-important predictors, i.e., *rent*, *location* and *house type*. For this, we employ the ensemble class probability p. Here, ensemble class probability refers to the predicted probability for a given class for a covariate

Table 2. Comparison of predictor importance.

Model	Excluding household attributes		Including household attributes	
Predictor	Mean decrease of accuracy[a]	Mean minimal depth[b]	Mean decrease of accuracy[a]	Mean minimal depth[b]
Rent	1.000 [1][c]	0.775 [1]	1.000 [1]	1.037 [1]
Location	0.580 [2]	1.342 [3]	0.610 [3]	1.600 [3]
House type	0.560 [3]	1.220 [2]	0.690 [2]	1.365 [2]
Rooms	0.430 [4]	2.470 [5]	0.410 [4]	2.815 [7]
Total area	0.400 [5]	2.021 [4]	0.380 [6]	2.658 [4]
Neighbourhood amenities	0.260 [6]	2.585 [6]	0.210 [10]	2.817 [8]
Furnishing	0.160 [7]	2.703 [7]	0.180 [11]	3.065 [9]
Income	–	–	0.400 [5]	2.702 [5]
Qualification	–	–	0.340 [7]	2.778 [6]
Age	–	–	0.290 [8]	3.300 [10]
Employment status	–	–	0.260 [9]	3.690 [11]

[a] RandomForest model, importance given as relative score in relation to most important predictor.
[b] RandomForestSRC model.
[c] Numbers in squared brackets indicate the rank of the variable, with 1 being equal to most important.

of interest $X = x_{i,}$ and with the values of all other predictors taken as-is. Thereby, the variance of predicted probabilities can be visualized, e.g., to identify factors that likely result in positive or negative residential choices.

Figure 1 shows this variance of p for the prediction of the minority class, p_{yes}. Looking at Figure 1(a) left, lower rents are apparently favoured, as indicated by the high median $\tilde{p}_{yes}$ of p_{yes}. E.g., for rents $\leq$ 300 EUR, $\tilde{p}_{yes} = 90.7\%$. I.e., for half the cases of the subset featuring this rent, the RF-predicted likelihood of a positive residential choice was greater than 90.7%, thus rendering these lowest rents a likely 'pull factor'. In contrast, for rents >1200 EUR, that median probability is only 7.8%.

Apparently, the interquartile range (IQR) differs substantially across factors. A higher IQR indicates a greater variance of p_{yes}, which is attributed to the variation of domain values of the remaining predictors. E.g., for *rent*, it is suggested that whilst rents of 300–700 EUR are generally favoured, other factors may nonetheless lead to a rejection of these scenarios. Rents of 700–1200 EUR have been deemed less favoured overall. However, the remaining factors obviously influence a participants' 'willingness-to-pay', possibly resulting in an acceptance of higher rents. Such rents could thus be seen as 'negotiable'. Regarding variable importance, *location* and *house type* might then be the most-important follow-up predictors governing these 'negotiations'. Limits to this 'willingness-to-pay' are reached by rents >1200 EUR, as indicated by the low $\tilde{p}_{yes}$ in conjunction with the small IQR, which could thus be considered as a 'deterring factor'. The ensemble class probabilities for the second and third-most important predictor can be interpreted accordingly. Clearly, the housing estate Grünau, the urban hinterland and the outskirts are the least favoured locations (Figure 1(a), middle). These findings confirm previously conducted resident surveys (Stadt Leipzig et al., 2016). For the other locations, $\tilde{p}_{yes}$ is comparably high, but so is their IQR. This is again interpreted as locations are not chosen at all costs. E.g., high rents or specific house types may be deterring factors. Looking at *house type* (Figure 1(a), right), housing estates from the 1970s and 1980s are less favoured. However, the IQR is comparatively high across all house types, which is again seen as the influence of the remaining predictors.

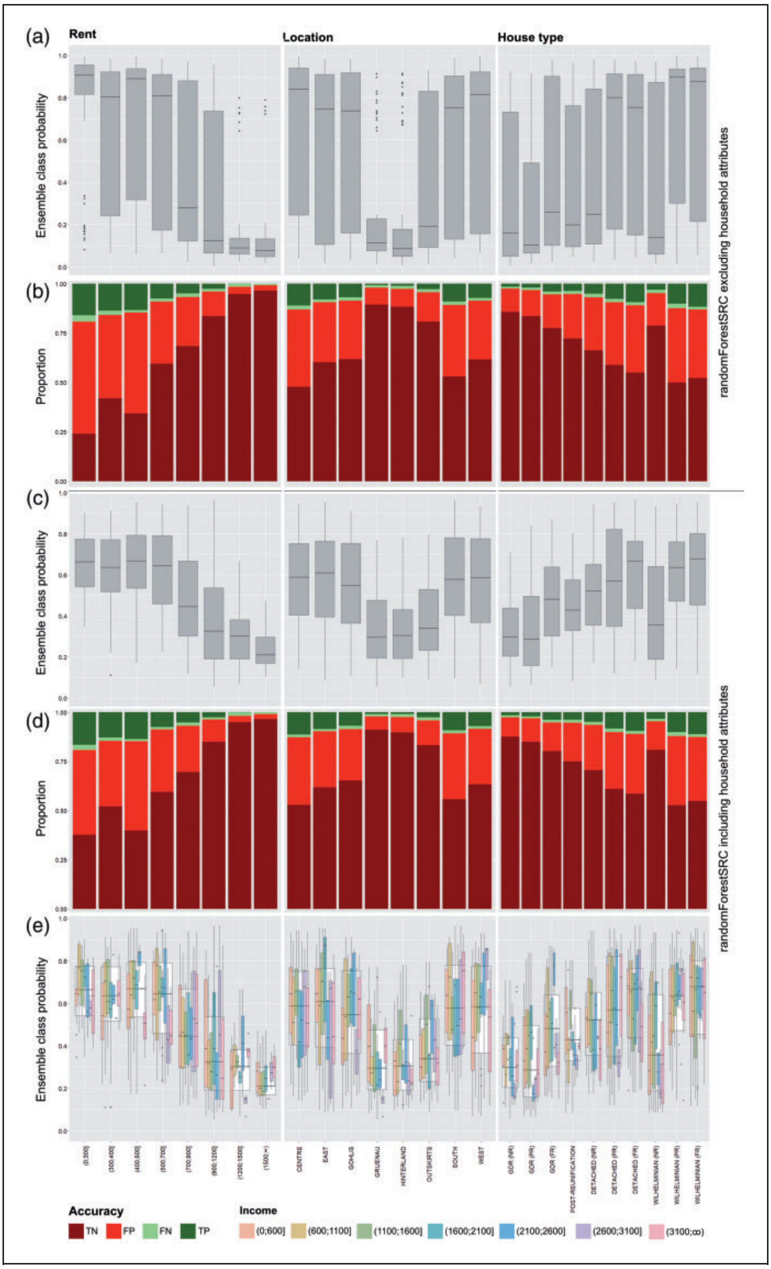

Figure 1. Variance of predicted ensemble class probabilities for the minority class for the predictors rent (left), location (location), and house type (right), and corresponding uncertainty, indicated by the proportions of TN and FP (positive reference outcomes), and FN and TP (negative reference outcomes). (a) Ensemble class probability of the randomForestSRC model excluding household attributes. (b) Uncertainty of randomForestSRC model excluding household attributes. (c) Ensemble class probability of randomForestSRC model including household attributes. (d) Uncertainty of randomForestSRC model including household attributes. (e) Ensemble class probability of randomForestSRC model including household attributes, overlaid with the ensemble probabilities broken down on groups of income as most-important sociodemographic predictor. TN: true negative; FP: false positive; FN: false negative; TP: true positive; NR: not renovated; PR: partially renovated; FR: fully renovated.

Here, possible interaction effects, i.e., the effect of a given predictor level on the preference of another, become apparent, e.g., (un-)preferred house types being associated with certain (un-)preferred locations. Grünau, e.g., is a large German Democratic Republic (GDR) housing estate built during the 1970s and 1980s, with a negative image (Grossmann et al., 2015).

High IQR could also indicate heterogeneous decision-making, where predictors have a strong influence on the individual decision outcomes, and where not only housing and neighbourhood attributes interact with each other, but where residential choice is further underpinned by household attributes. Hence, including household attributes as additional predictors may improve the predictive performance by decreasing the variance of p. Figure 1(c) visualizes the respective ensemble class probabilities for the corresponding randomForestSRC model. For most factors, IQR has been reduced, and the median probabilities $\tilde{p}_{yes}$ were adjusted accordingly. Ensemble class probabilities can be broken down further using household characteristics. Figure 1(e) exemplifies this using the most-important household attribute, *income*. Looking at Figure 1(e), it becomes clear how p_{yes} varies between different classes of household income.

Model accuracy and uncertainty analysis

Heretofore, the importance of predictors has been elicited, and, for the three most-important variables *rent*, *location* and *house type*, it has been discussed how the probability of predicting positive residential choices varies across factors. In the following, the accuracy of these predictions will be evaluated using the OOB error as an internal measure of RF accuracy (Breiman, 2001; Hastie et al., 2009). Additional measures will be derived from confusion matrices, which compare predictions with known responses of the test set. Thereby, the number of true- and false-negative predictions (TN, FN), as well as true and false-positive outcomes (TP, FP) are determined (Rokach and Maimon, 2015). Based on these numbers, the following accuracy measures have been computed (Table 3): (i) success rate, i.e., the fraction of correct predictions from all observations; (ii) recall, i.e., the fraction of correctly identified TP from all positive reference outcomes; (iii) specificity, i.e., the proportion of TN correctly identified as such and (iv) precision, i.e., the proportion of TP from all predicted positive outcomes (Witten et al., 2011).

Looking at Table 3, the OOB error is similar across all RF models. This is also true for the success rate and specificity. Recall is similar to the success rate for models excluding household attributes, but slightly higher for models including them. Precision is comparatively low for all models, which indicates that TP are contrasted by a

Table 3. Comparison of random forest (RF) accuracy measures.

	Excluding household attributes		Including household attributes	
Measure (%)	Random-Forest	Random-ForestSRC	Random-Forest	Random-ForestSRC
OOB error (minority class)	21.40	21.80	21.91	21.91
Success rate	74.32	75.23	77.80	78.21
Recall	77.31	73.95	84.55	84.55
Specificity	74.07	75.33	77.38	77.61
Precision	19.91	20.00	26.07	26.26

OOB: out-of-bag error.

considerable number of FP, giving rise to uncertainty regarding positive predictions being made. Similar to recall, including household attributes slightly improves the model performance.

The previously introduced accuracy measures are subsequently also determined on a per-factor basis to uncover factor levels for which uncertainty is particularly high. This is done by recording TP, FP, TN and FN for each covariate of interest, using 100 iteratively grown randomForestSRC models. Figure 1 visualizes the resulting mean fractions of TP and FN – both corresponding to true positive reference outcomes in the test sets – as well as TN and FP, both representing true negative reference outcomes, for the most-important predictors.

As discussed above, uncertainty results from false predictions, i.e., FP and FN. In Figure 1, FPs are shown in light red, and FN in light green. Looking at Figure 1(b) and 1(d), the share of FN is comparatively low across all factors and models, thus rendering FN the minor contributor to uncertainty. The number of FP is clearly higher, especially for lower rents, locations in the Centre and the South of Leipzig, and fully-renovated detached as well as fully or partially-renovated Wilhelminian-style houses. This is particularly true for RF models not considering household attributes as shown in Figure 1(b). In comparison, looking at Figure 1(d), models including household attributes seem to perform better especially for lower rents and central location. For these factors, the number of FP has clearly decreased.

Discussion and conclusions

Technical-methodical KnETs advancement

This paper has showcased an advanced KnETs approach called SilverKnETs for residential choice modelling in the city of Leipzig, Germany. The SilverKnETs tool as a technical KnETs advancement has proven to be a robust data collection instrument. Using RF as a methodological KnETs advancement, we were able to predict residential choice and assess predictor importance based on either MDA or the mean minimal depth. Using ensemble class probabilities, we were further able to uncover factors that likely lead to positive ('pull') or negative ('push') residential choices as well as identifying likely interactions between these factors. Hence, the approach allows dependencies to be detected between variables, e.g., under which conditions factors such as high rents or disadvantaged locations become more acceptable.

However, we believe that not only those conditions that are perceived as attractive, are of relevance to planners, but also that knowledge about conditions that make residential locations undesirable, thus reinforcing their unattractiveness, is valuable information for stakeholders (Stadt Leipzig et al., 2016). The identification of those factors deemed as (un-)preferred and (un-)important could provide valuable insights in how to develop the housing market to cater to public demand, e.g., in the context of neighbourhood developments and housing stock planning.

Clearly, the proposed KnETs adaptation is not a completely new methodology *per se*, but is meant to complement established methodologies such as choice-based conjoint analysis (CBC), an increasingly popular method for the elicitation of preferences in environmental and conservation issues (Alriksson and Öberg, 2008). Consequently, there are many similarities. Both methods can reveal the preference structure of individuals or groups. Likewise, also CBC is able to reveal the (relative) importance of attributes and predictors (Alriksson and Öberg, 2008; Brett Hauber et al., 2016). However, a very clear and considerable advantage of tree-based methods is their ability to visualize the hierarchical structure of decision-making, i.e., the rules that a decision follows on a step-by-step basis (Figure S3).

Revisiting predictors and preference heterogeneity

This case study has been based on a limited, expert-elicited set of explanatory variables. From this set of variables, *rent*, *location* and *house type* were consistently identified as the most-important predictors across all RF models. This finding confirms questionnaire surveys undertaken in Leipzig by Welz et al. (2014, 2017), which concluded that housing and neighbourhood attributes appear to be more important than household attributes. Regarding the remaining predictors, there is no clear-cut order of importance. Following Strobl et al. (2008), this may also be due to interactions and complex correlations, which may affect the estimation of variable importance (Hothorn et al., 2006).

Regarding the re-evaluation of predictors, transport-related criteria, e.g., distance to city centre, specific institutions or well-known landmarks could possibly be included. Heterogeneity in preferences in residential choice should additionally be considered (Ettema, 2010; Walker and Li, 2007). It could be shown that by including household attributes, the variance of predicted ensemble class probabilities could be reduced, consequently increasing the predictive performance of the RF models. However, looking at Table 3, the observed increase in precision is surprisingly small and possibly lower than anticipated. Particularly recall seems to have benefitted from including household attributes. I.e., the models were more successful in identifying TP.

In the context of the presented case study, particularly if there is a focus on predicting rather than explaining residential choice, the elicitation of household attributes might thus be re-considered. Instead, predictors could be elicited that reflect on past experiences and choices as well as shocks or changes during the household-aging process. As discussed by Bajari et al. (2013), Clark and Huang (2003) or Yu et al. (2017), these aspects may have significant influences on present and future preferences and behaviour.

Nevertheless, household attributes allow for a further differentiation of predicted class probabilities. As exemplified in Figure 1(e), ensemble class probabilities vary – for some cases considerably – per category of *income*. However, overall, *income* is not a very important predictor, and consequently, it does not seem to dominate residential choice. We attribute this to the fact that the housing market in Leipzig at the time of the survey was less tense and polarized and thus more accessible to most households compared to, e.g., London or Paris. We expect that a stronger polarization of the housing market would clearly be reflected in the importance ranking of the predictors, which could be a suitable avenue for further research. As discussed by Walker and Li (2007), lifestyles might be another crucial factor for the explanation of some of the observed patterns. E.g., looking at the median ensemble class probabilities shown in Figure 1(e), left, it becomes clear that the highest rent class is ‘preferred’ to almost the same extent in both the lowest and the highest category of income. For the former income category, this might be explained by lifestyle choices such as flat sharing.

Reflections on the performance and limitations of RF

From the assessment of the RF models, it became obvious that neither model is superior. We have shown that the precision of predicting the minority class is, depending on the model, rather low with 20–26%. This is mainly due to a high number of FP. It appears to be the case that RFs tend to overestimate the impact of favourable conditions such as low rent, resulting in overly ‘optimistic’ predictions and thus rather high uncertainty. To decrease this uncertainty, the number of FP could be reduced by (i) penalizing false predictions; (ii) re-evaluating predictor selection; and (iii) increasing sample size.

Comparing the elicited RF models to a binary logistic regression (Table S4), it can be concluded that the precision of both types of models is of comparable magnitude, but that recall is higher for RF (Table S5). This also applies to a comparison with CART, where the precision of RF is only slightly higher, but recall considerably so (Table S6). It is noteworthy that there is a trade-off between recall and precision, where maximizing any of the two measures results in a degrading performance of the other (Rokach and Maimon, 2015). Hence, highly precise models tend to lack in detecting positive decisions. With increasing recall, on the contrary, precision decreases. This limiting trade-off requires optimization of a model on a case-by-case basis.

Transferability and re-use of SilverKnETs

The generic nature of SilverKnETs allows for a flexible transfer and re-use of the tool across problem domains and use cases. The approach presented here can generally be used within any domain where stakeholder preferences in a decision-making context need to be elicited. This includes, e.g., resource management (Price et al., 2016) or land-use planning (Vollmer et al., 2016). The different knowledge elicitation games included in SilverKnETs, the integration of a dedicated view model that determines how information is presented to the participants on-screen – e.g., regarding language or level of detail – as well as the flexibility offered by the tool to conduct interviews offline or online are seen to facilitate this reusability and transferability.

Declaration of conflicting interests

The author(s) declare no potential conflict of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Sebastian, Dagmar and Manuel benefited from funding through the FP7 collaborative project GREENSURGE (FP7-ENV.2013.6.2-5-603567), the Horizon 2020 innovation action CONNECTING (COproductioN with NaturE for City Transitioning, Innovation and Governance; No 730222-2) and the AXA research award 'Models, Metrics and Typologies of Resilient Cities'. Nadja's work was supported by Green-EquityHEALTH, funded by the German Federal Ministry of Education and Research (BMBF; no. 01LN1705A). This research was carried out as part of the project ENABLE, funded through the 2015-2016 BiodivERsA COFUND call for research proposals, with the national funders The Swedish Research Council for Environment, Agricultural Sciences, and Spatial Planning, Swedish Environmental Protection Agency, German aeronautics and space research centre, National Science Centre (Poland), The Research Council of Norway and the Spanish Ministry of Economy and Competitiveness.

References

Alriksson S and Öberg T (2008) Conjoint analysis: A useful tool for assessing preferences for environmental issues. *Environmental Science and Pollution Research* 15(2): 119.

Angelini V and Laferrère A (2012) Residential mobility of the European elderly. *CESifo Economic Studies* 58(3): 544–569.

Antipov EA and Pokryshevskaya EB (2012) Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications* 39(2): 1772–1778.

Bajari P, Chan P, Krueger D, et al. (2013) A dynamic model of housing demand: Estimation and policy implications. *International Economic Review* 54(2): 409–442.

Bartkowski B, Lienhoop N and Hansjürgens B (2015) Capturing the complexity of biodiversity: A critical review of economic valuation studies of biological diversity. *Ecological Economics* 113: 1–14.

Belgiu M and Drăguţ L (2016) Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114: 24–31.

Bharwani S (2006) Understanding complex behavior and decision making using ethnographic knowledge elicitation tools (KnETs). *Social Science Computer Review* 24(1): 78–105.

Bharwani S, Bithell M, Downing TE, et al. (2005) Multi-agent modelling of climate outlooks and food security on a community garden scheme in Limpopo, South Africa. *Philosophical Transactions of the Royal Society B – Biological Sciences* 360(1463): 2183–2194.

Bharwani S, Coll Besa M, Taylor R, et al. (2015) Identifying salient drivers of livelihood decision-making in the forest communities of Cameroon: Adding value to social simulations. *Journal of Artificial Societies and Social Simulation* 18(1): 3.

Bhat CR and Guo J (2004) A mixed spatially correlated logit model: Formulation and application to residential choice modeling. *Transportation Research Part B* 38: 147–168.

Breiman L (2001) Random forests. *Machine Learning* 45: 5–32.

Brett Hauber A, González J, Groothuis-Oudshoorn C, et al. (2016) Statistical methods for the analysis of discrete choice experiments: A report of the ISPOR conjoint analysis good research practices task force. *Value in Health* 19(4): 300–315.

Chen C, Liaw A and Breiman L (2004) *Using Random Forest to Learn Imbalanced Data*. Report, Report no. 666, July. Berkeley: University of California.

Clark WAV and Huang Y (2003) The life course and residential mobility in British housing markets. *Environment and Planning A* 35: 323–339.

Einav L and Levin J (2014) Economics in the age of big data. *Science* 346(6210): 1243089.

Ettema D (2010) The impact of telecommuting on residential relocation and residential preferences. *Journal of Transport and Land Use* 3(1): 7–24.

Grossmann K, Kabisch N and Kabisch S (2015) Understanding the social development of a post-socialist large housing estate: The case of Leipzig-Grünau in eastern Germany in long-term perspective. *European Urban and Regional Studies*. 24(2): 142–161.

Gustafsson A, Herrmann A and Huber F (2001) Conjoint analysis as an instrument of market research practice. In: Gustafsson A, Herrmann A and Huber F (eds) *Conjoint Measurement – Methods and Applications*. Berlin: Springer, pp. 3–30.

Haase A, Herfert G, Kabisch S, et al. (2012a) Reurbanizing Leipzig (Germany): Context conditions and residential actors (2000–2007). *European Planning Studies* 20(7): 1173–1196.

Haase D, Kabisch N, Haase A, et al. (2012b) Actors and factors in land use simulation – the challenge of urban shrinkage. *Environmental Modelling and Software* 35: 92–103.

Haase D, Lautenbach S and Seppelt R (2010) Applying social science concepts: Modelling and simulating residential mobility in a shrinking city. *Environmental Modelling and Software* 25: 1225–1240.

Haase A and Rink D (2015) Inner-city transformation between reurbanization and gentrification: Leipzig, eastern Germany. *Geografie* 15(2): 226–250.

Hastie T, Tibshirani R and Friedman J (2009) *The Elements of Statistical Learning*, 2nd ed. New York: Springer.

Hothorn T, Hornik K and Zeileis A (2006) Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3): 651–674.

Ibraimovic T and Masiero L (2012) Do birds of a feather flock together? The impact of ethnic segregation preferences on neighbourhood choice. *Urban Studies* 51(4): 693–711.

Ishwaran H, Kogalur UB, Blackstone EH, et al. (2008) Random survival forests. *The Annals of Applied Statistics* 2(3): 841–860.

Ishwaran H, Kogalur UB, Chen X, et al. (2011) Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining* 4: 115–132.

Kabisch N, Haase D and Haase A (2010) Evolving reurbanisation? Spatio-temporal dynamics exemplified at the eastern German city of Leipzig. *Urban Studies* 47(5): 967–990.

Kemp-Benedict EJ, Bharwani S and Fischer MD (2010) Methods for linking social and physical analysis for sustainability planning. *Ecology and Society* 15(3): 4.

Kim J-H, Pagliara F and Preston J (2003) An analysis of residential location choice behaviour in Oxfordshire, UK: A combined stated preference approach. *International Review of Public Administration* 8(1): 103–114.

Lauf S, Haase D, Seppelt R, et al. (2012) Simulating demography and housing demand in an urban region under scenarios of growth and shrinkage. *Environment and Planning B* 39: 229–246.

Lausch A, Schmidt A and Tischendorf L (2015) Data mining and linked open data – New perspectives for data analysis in environmental research. *Use of Ecological Indicators in Models* 295: 5–17.

Liaw A and Wiener M (2002) Classification and regression by randomForest. *R News* 2(3): 18–22.

López-Ospina HA, Martínez FJ and Cortés CE (2016) Microeconomic model of residential location incorporating life cycle and social expectations. *Computers, Environment and Urban Systems* 55: 33–43.

McFadden D (1978) Modeling the choice of residential location. *Transportation Research Record* 673: 72–77.

Park J and Kim K (2016) The residential location choice of the elderly in Korea: A multilevel logit model. *Journal of Rural Studies* 44: 261–271.

Price JI, Janmaat J, Sugden F, et al. (2016) Water storage systems and preference heterogeneity in water-scarce environments: A choice experiment in Nepal's Koshi River Basin. *Water Resources and Economics* 13: 6–18.

Raymond CM, Fazey I, Reed MS, et al. (2010) Integrating local and scientific knowledge for environment management. *Journal of Environmental Management* 91: 1766–1777.

Rodriguez-Galiano VF, Ghimire B, Rogan J, et al. (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 67: 93–104.

Rokach L and Maimon O (2015) *Data Mining With Decision Trees: Theory and Applications*, 2nd ed. Singapore: World Scientific.

Sener IN, Pendyala RM and Bhat CR (2011) Accommodating spatial correlation across choice alternatives in discrete choice models: An application to modeling residential location choice behavior. *Journal of Transport Geography* 19: 294–303.

Stadt Leipzig, Amt für Wahlen und Statistik (2016) *Kommunale Bürgerumfrage 2015*. Report, Leipzig, June. Leipzig: Stadt Leipzig, Amt für Statistik und Wahlen.

Stokenberga A (2019) How family networks drive residential location choices: Evidence from a stated preference field experiment in Bogotá, Colombia. *Urban Studies* 56(2): 368–384.

Strobl C, Boulesteix A-L, Kneib T, et al. (2008) Conditional variable importance for random forests. *BMC Bioinformatics* 9: 307.

Timmermans H, Borgers A, van Dijk J, et al. (1992) Residential choice behaviour of dual earner households: A decompositional joint choice model. *Environment and Planning A* 24: 517–533.

Tu G, Abildtrup J and Garcia S (2016) Preferences for urban green spaces and peri-urban forests: An analysis of stated residential choices. *Landscape and Urban Planning* 148: 120–131.

Vollmer D, Ryffel AN, Djaja K, et al. (2016) Examining demand for urban river rehabilitation in Indonesia: Insights from a spatially explicit discrete choice experiment. *Land Use Policy* 57: 514–525.

Walker JL and Li J (2007) Latent lifestyle preferences and household location decisions. *Journal of Geographical Systems* 9(1): 77–101.

Welz J, Haase A and Kabisch S (2014) Meine Entscheidung für Leipzig. Ergebnisse der Wanderungsbefragung 2014. In: Leipzig Stadt and für Wahlen und Statistik Amt (eds) *Statistischer Quartalsbericht II/2014*. Leipzig: Stadt Leipzig, pp. 19–24.

Welz J, Haase A and Kabisch S (2017) Zuzugsmagnet Grossstadt – Profile aktueller Zuwanderer. *disP – The Planning Review* 53(3): 18–32.

Witten IH, Frank E and Hall MA (2011) *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington: Elsevier.

Wolff M and Haase A (2015) Stadtregion Leipzig-Halle jenseits der Schrumpfung: neues Wachstum und Stabilisierung. In: Leipzig Stadt and für Wahlen und Statistik Amt (eds) *Statistischer Quartalsbericht I/2015*. Leipzig: Stadt Leipzig, pp. 36–42.

Wolff M, Haase A, Haase D, et al. (2016) The impact of urban regrowth on the built environment. *Urban Studies*. 54(12): 2683–2700.

Wood LE and Ford JM (1993) Structuring interviews with experts during knowledge elicitation. *International Journal of Intelligent Systems* 8(1): 71–90.

Wooten TC and Rowley TH (1995) Using anthropological interview strategies to enhance knowledge acquisition. *Expert Systems with Applications* 9(4): 469–482.

Wright MN, Ziegler A and König IR (2016) Do little interactions get lost in dark random forests? *BMC BioInformatics* 17: 145.

Yates J and Mackay DF (2006) Discrete choice modelling of urban housing markets: A critical review and an application. *Urban Studies* 43(3): 559–581.

Yu B, Zhang J and Li X (2017) Dynamic life course analysis on residential location choice. *Transportation Research A: Policy and Practice* 104: 281–292.

Sebastian Scheuer holds a diploma from Martin Luther University Halle-Wittenberg and a PhD from Humboldt-Universitt zu Berlin, where he currently is a Postdoctoral Researcher at the Landscape Ecology Lab. His research is centred on human-environmental interactions and includes the knowledge-driven, holistic and integrative multi-criteria assessment of natural hazard and climate change-related vulnerabilities and risks for the urban space from a local, regional, and global perspective. He has worked on the elicitation and formalization of knowledge using data-mining and ontologies, and on the implementation of knowledge-based systems. Recently, he focused on the qualitative and quantitative, spatially explicit modelling of processes of urbanization at different spatial scale.

Dagmar Haase is a Professor in urban ecology and urban land use modelling. She holds a PhD from the Martin-Luther-University of Halle-Wittenberg. Dagmar is professor at the Humboldt Universität zu Berlin, Germany, and Guest Scientist at the Helmholtz Centre for Environmental Research – UFZ. Dagmar's main interest and activities are settled in the combination and integration of global urbanization modelling and the quantification and assessment of ecosystem services, disservices, green infrastructure and social-environmental justice issues in cities and urban areas including urban land teleconnections. She works at different spatial scales, from the global to the local and neighbourhood scale. Conceptually, Dagmar's Lab bases its work on the idea of emergence, resilience and sustainability of social-ecological coupled systems. Geographically, her focus areas are situated in Europe and in Russia. She is author of over 150 ISI-listed scientific publications. In 2010, Dagmar was Fellow of the International Environmental Modelling & Software Society (iEMSs), in 2014, she received the AXA Award for research on ''Resilient Cities''. In 2016, Dagmar held the Honorary Wallenberg Professorship of the Swedish Academy of Sciences.

Annegret Haase is a Postdoctoral Researcher at the Department of Urban and Environmental Sociology at the Helmholtz Centre for Environmental Research – UFZ in Leipzig, Germany. Her main research interests include sustainable urban transformation, urban shrinkage and reurbanization, land use change in cities, urban green areas and urban ecosystem services, socio-spatial processes in cities, inequalities and urban diversity, urban governance and postsocialist cities. She has been involved in numerous EU projects (FP5, FP7) since 2002 including the coordination of an SSH-2007-2.2.1 project Shrink Smart (grant agreement no. 225193, 2009-2012) and UFZ team coordination in an SSH.2012.2.2-1 project Divercities (Grant agreement no: 319970, 2009-2013). Currently, she is involved in two projects (KoopLab, MigraChance) funded by the German national ministry of research (BMBF) which deal with the collaborative shaping of green spaces in arrival neighbourhoods and the analysis of migration-related conflicts as triggers for institutional learning. Her research is strongly interdisciplinary oriented and includes international comparative studies.

Nadja Kabisch holds a PhD and a Diploma in Human Geography from the Martin-Luther University Halle-Wittenberg, Germany, and is a Senior Researcher at the Humboldt-Universitt zu Berlin in the Department of Geography, Landscape Ecology Lab. Nadja has worked in the EU FP7 project GREENSURGE and the BiodivERsA projects URBES and ENABLE. Her special interest is on human-environment interactions in cities taking co-benefits from nature-based solutions implementation for human health and social justice into account. Since August 2017 she leads a 5-year junior research group "GreenEquityHEALTH" on the interlinkages between urban green spaces, public health and socio-environmental justice in European cities.

Manuel Wolff is a Research Fellow at the Humboldt-Universität zu Berlin in the Department of Geography, Landscape Ecology Lab. He has worked in the integrated project 'Urban Transformations' at the Helmholtz-Centre for Environmental Research (UFZ), in the networking project 'CIRES – Cities Regrowing Smaller' and other projects at the Université Paris Sorbonne, University of Nottingham or Charles University Prague. Manuel's main interest is settled in land use changes within the context of urban growth and shrinkage and the associated impacts on ecosystem services. Especially, he works on quantitative comparative analysis of social-ecological structures and trends in European cities including the development of indicator concepts and GIS-modelling.

Nina Schwarz is an Assistant Professor at the University of Twente, Faculty of Geo-Information Science and Earth Observation – ITC. She holds a diploma in environmental sciences from the University of Lueneburg and a PhD in social and economic sciences from the University of Kassel. After her PhD she worked at the Helmholtz Centre for Environmental Research – UFZ in Leipzig. Her research interests lie in interdisciplinary questions on urban development, specifically urban land use change and urban ecology. For her project, Nina uses methods of modelling and simulation, such as agent-based modelling, and quantitative statistics.

Katrin Großmann is a Professor in Urban Sociology at University of Applied Sciences Erfurt since 2014. In her research, she combines different aspects of justice at the intersections of critical urban studies, sociological theory and environmental topics such as energy poverty, energy policy and housing; social cohesion, migration and conflict; social inequality, residential segregation and neighbourhoud change. Urban shrinkage has been a long term interest in her work which lately lead to a focus on small cities and their trajectories. From a methodological point of view, her expertise is in mixed method approaches.