In free float: Developing Business Analytics support for carsharing providers☆

Sebastian Wagner, Tobias Brandt*, Dirk Neumann

Information Systems Research, University of Freiburg, Platz der Alten Synagoge, 79098 Freiburg, Germany

**A R T I C L E   I N F O**

Article history:
Received 15 July 2014
Accepted 19 February 2015
Available online 30 March 2015

Keywords:
Free-floating carsharing  
Decision support  
Data analytics  
Spatial analysis

**A B S T R A C T**

As a rapidly expanding market, carsharing presents a possible remedy for traffic congestion in urban centers. Especially free-floating carsharing, which allows customers to leave their car anywhere within the operator’s business area, provides users with flexibility, and complements public transportation. We present a novel method that provides strategic and operational decision support to companies maneuvering this competitive and constantly changing market environment. Using an extensive set of customer data in a zero-inflated regression model, we explain spatial variation in carsharing activity through the proximity of particular points of interests, such as movie theaters and airports. As an application case, as well as a validation of the model, we use the resulting indicators to predict the number of rentals before an expansion of the business area and compare it to the actual demand post-expansion. We find that our approach correctly identifies areas with a high carsharing activity and can be easily adapted to other cities.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

In the past decade, carsharing as a business model has experienced a tremendous growth. In North America alone, carsharing membership has increased at a compound annual growth rate exceeding 45 percent, from 16,000 in 2002 to more than a million members in early 2013 [1]. Yet, Shaheen and Cohen [2] emphasize that the growing popularity of carsharing is a global phenomenon, with an established market in Europe and increasing interest in Asia and Australia. Furthermore, it is considered to be a possible cure to ailments urban centers around the world are suffering from, such as greenhouse gas emissions, air pollution, and traffic congestion [3]. Carsharing relies on synergies with public transit – busses and trains serve as a reliable infrastructure for daily commutes, with carsharing providing additional flexibility.

This flexibility is maximized by the concept of free-floating carsharing (FFC), which disposes of static rental stations and allows customers to end their rentals anywhere within the carsharing provider’s operating area [4]. Carsharing members use smartphone applications to locate and reserve the closest available vehicle, RFID cards provide access to the vehicle, and an online payment system handles the billing. However, for the operator FFC concepts are associated with substantially increased complexity – both, with respect to strategic as well as operational decisions. On the strategic level, free-floating carsharing providers need to define their operation area by deciding whether certain parts of the city should be left out, and by determining optimal expansion strategies. These decisions directly translate to the operational level. Vehicle demand and supply vary during the day and across locations, such that operators have to relocate vehicles within the operation area in order to provide a satisfactory service level to members.

In this paper we address the challenges outlined in the previous paragraph, namely the strategic and operational issues arising from a temporal and spatial divergence between vehicle demand and supply, and demonstrate the substantial value Business Analytics methods and applications can bring to emerging business sectors [5,6] such as carsharing. We introduce an innovative analytics tool that aims to relate the spatial variation of vehicle demand to points of interest (POIs) in the vicinity, which include, for instance, bus stops, movie theaters, and shopping malls. Through the cooperation with a globally leading carsharing provider, our analysis is based on a dataset spanning more than a million carsharing rentals in the city of Berlin between April 2012 and October 2013. Employing a zero-inflated Poisson model, we relate these observations to more than 180,000 POIs, as well as demographic data, in order to identify POI categories that substantially influence variation in carsharing activity. Furthermore, we illustrate the value of our approach by providing strategic decision support on possible expansion areas, which is validated by comparing the predicted with actual vehicle demand after an expansion of the operating area occurred in 2013.
Consequently, the hypotheses investigated during the course of this paper are the following:

**Hypothesis 1.** Points of interest, as a proxy for the attractiveness of the various areas within a city, can be used to explain the spatial variation in the demand for carsharing.

**Hypothesis 2.** This relationship can be used to provide decision support regarding expansion strategies.

The paper is structured as follows. In Section 2, we discuss research related to our work. In Section 3, we outline the theoretical argument for a statistical relationship between POIs and carsharing activity, and describe our dataset. In Section 4, the regression analysis is conducted, investigating Hypothesis 1. In Section 5, we analyze the applicability of our results to an expansion of the operation area, which relates to Hypothesis 2. Section 6 concludes.

### 2. Related work and contribution

In this section, we first present an overview on carsharing and related research. This is followed by a brief summary of geospatial research as it pertains to our work. We conclude this section by positioning our research in a Business Analytics context and specifying our contribution.

#### 2.1. Carsharing

The history of carsharing dates back to 1948 and the SEFAGE ("Selbstfahrgemeinschaft") in Zurich, Switzerland [7]. However, it took until the late 1980s and early 1990s for carsharing to become more common, particularly in central Europe, as well as Great Britain, Italy, and the Scandinavian countries [8]. In the United States carsharing gained popularity following several pilot projects, which aimed to better understand how carsharing can be established and operated [9]. Traditionally, carsharing operators have provided station-based services. Round-trip systems require the customer to return the car to the station where it was originally picked up, while one-way systems afford more flexibility by allowing customers to end their rentals at any station designated by the provider. The novel free-floating concept takes this increased flexibility one step further by allowing customers to return vehicles anywhere within a certain operation area.

Naturally, this evolution in carsharing concepts has been accompanied by a change in the research objectives related to carsharing. Questions regarding the (in)activity of users [10], the lengths and durations of trips [11], and the spatial distribution and determinants of carsharing demand [12,13] concern all carsharing systems. Moreover, the introduction of one-way and free-floating concepts have raised a range of new issues related to the temporal and spatial imbalance between supply of and demand for vehicles. This problem becomes even more pronounced if we take the changing needs of electric vehicles into account [14]. On the other hand, sharing a vehicle can provide benefits from a social and environmental perspective by reducing the overall number of cars on the road [15]. Designing an optimal sharing system regardless of the means of transport remains difficult for station based services [16] and even more difficult in the free-floating model.

Relocation mechanisms that seek to alleviate this imbalance have been investigated by several research groups [17,18]. For instance, Barth et al. [19] introduce a user-based relocation mechanism that encourages users to share or split rides depending on system-wide demand. Kek et al. present event-based relocation techniques validated by real-world data [20]. They implement a complete decision support system that was tested with a carsharing company in Singapore [21], a city that has been at the forefront of rethinking car ownership for more than two decades [22]. The problem of relocating vehicles can also be found in the bike sharing domain. Dell’Amico et al. [23] provide a linear mixed integer programming formulation with the objective to minimize total costs for bike sharing. For FFC systems, Weikl and Bogenberger [4] introduce a relocation system consisting of an offline module that pre-calculates possible relocation strategies based on historical data and an online module, which adjusts these strategies according to the current situation. Critical success factors for carsharing systems have been researched by Millard-Ball [24], as well as Celsor and Millard-Ball [25]. They find that the purposes of rentals are often associated with specific points of interest, such as grocery stores and shops. However, the actual relationship between POIs and carsharing demand has not yet been investigated. Since such an empirical model would provide valuable input to relocation mechanisms, as well as long-term strategic decision-making, the first hypothesis guiding this paper addresses this issue.

#### 2.2. Geospatial research

The emergence of location-based services [26–28] as a growing research field has introduced geospatial analysis [29,30] into the methodological toolbox of Information Systems and Management researchers. One increasingly used method, the geographically-weighed regression (GWR) [31], explores nonstationarity in geographic regression parameters. It calculates local coefficients by distance-weighing observations. This method is purely explanatory and difficult to use for prediction. Hence, it is of limited use in our context, given that one application of our approach is to estimate carsharing activity in locations without any existing carsharing data (expansions). Furthermore, Páez et al. [32] note the flaw of the GWR in its tendency towards an increased amount of false-positives, i.e. a faulty recognition of spatial nonstationarity. Considering these drawbacks, we refrain from using the GWR in favor of a zero-inflated count model (ZINF).

#### 2.3. Business Analytics

Business Analytics is one of the most prominent current research streams in Information Systems, as well as Management research [33,34]. It continues several decades of decision support research and the terms “Decision Support System”, “Business Intelligence”, and “Business Analytics” are often used interchangeably [35]. However, the core of Business Analytics applications is to unleash the potential of Big Data and substantially transform business operations through data-driven strategies [6]. The breadth of Business Analytics applications is illustrated by the variety of recent publications. Topics include the relationship between advertisements and purchasing decisions [36], resource allocation in emergency response scenarios [37], energy security in a changing world climate [38], as well as decision support for mergers and acquisitions [39].

In our work, we apply the transformative power of Business Analytics methods to carsharing – a business sector that has existed for several decades, but has been invigorated by mobile devices and data-driven business strategies such as FFC. Hence, the objective of our paper is not just to analyze carsharing data, but also to discuss how the insights were used for real-world business decisions. Building upon the second hypothesis guiding this paper, we demonstrate how the empirical results support strategic decision making regarding the expansion of the carsharing provider’s operation area, and outline further possible applications to relocation strategies.
2. Theoretical motivation and data basis

The theoretical motivation and data basis can provide operational guidance to carsharing operators. We also outline how our methodology can provide operational guidance to carsharing operators.

3. Theoretical motivation and data basis

In a provocative piece, Anderson [40] heralds the end of theory through data-driven research. Chang et al. [41] on the other hand outline the potential of ubiquitous data to better support or disprove existing theories, emphasizing the duality between theoretical and empirical research. In this paper, we follow this suggestion and employ novel data sources in conjunction with established theories. Our theoretical reasoning is based on the fact that most trips by car or other means of transportation are undertaken for purposes that can be associated with specific POIs, such as work, shopping, church, or recreational sports [42]. A crucial determinant of parking decisions, in turn, is the proximity of the prospective parking spot to the intended destination [43,44]. However, in a carsharing context, parking often corresponds with ending a rental, unless the anticipated time spent at the destination is short, since the customer pays for the vehicle during parking, as well. Hence, the vicinities of certain “attractive” POI categories should be characterized by high numbers of started and ended rentals. This relationship is further illustrated in Fig. 2. With a specific activity in mind, such as watching a newly released movie, the customer begins an FFC rental. The rental is ended when the car is parked close to the POI associated with the intended activity, i.e. the movie theater. The presence of the movie theater causes customers to move towards that particular location, affecting the demand for carsharing services in its vicinity.

The POIs in our analysis are derived from Google Maps and tagged with any combination of 92 categories. For instance, a specific restaurant can be tagged with restaurant, food, and meal takeaway. We investigate whether particular categories of POIs have a significant influence on the demand for carsharing vehicles in the surrounding areas. For example, a high density of restaurants (i.e. POIs tagged with the category restaurant) in a specific neighborhood may increase the demand for vehicles. The POI density – i.e. the number of POIs within the vicinity of specific location – is a core concept of our analysis and will be elaborated in more detail in the next section.

The data basis of our analysis contains information on POIs, carsharing use, and demographics for Berlin, Germany. We chose Berlin as a demonstration case for several reasons. First, by 2013 it had become the city with the largest one-way or free-floating carsharing system in the world [45], enabling a thorough analysis of carsharing behavior. Second, the provider’s operation area was expanded in November 2012. Since we analyze data from three months before and after this expansion, respectively, we are able to use it as a case study to validate our methodology. The data provided by the carsharing provider includes half a million reservations conducted by a customer base of approximately 55,000 members during the course of six months. These customers used 1200 shared vehicles within an operation area of almost 300 square kilometers. The destinations (GPS coordinates) of all rentals constitute the dependent variables of our regression model. In the course of this paper, we use rentals and trips as synonyms for reservations. While a single reservation can generally consist of multiple trips, we only analyze the starting and end points of the reservation and consider a reservation to be one trip. The explanatory variables are constructed from data on more than 180,000 POIs within the city of Berlin, as well as census data as additional control variables. In the next section, we will develop the empirical model designed to provide decision support based on this data set.

4. Empirical investigation of carsharing demand

In this section, we will first introduce the formal constructs, such as POI density, which transform the raw data into information that can be leveraged in the empirical analysis. Along with these definitions, we provide geospatial visualizations of the destinations of carsharing usage. This is followed by the actual regression analysis. We conclude this section by assessing our first hypothesis in light of the regression results.

4.1. Formal constructs and data visualization

To transform the rental data into meaningful input for the regression analysis and subsequently use the results for decision support, we first need to define the area of interest, i.e. the operating area of the carsharing provider within the city. For this purpose, we define a spatial area $A_i$ as a closed polygon of at least 3 different GPS-points, given by the tuple

$$A_i = (a_{i1}, \ldots, a_{in}, a_{i1})$$

4.1. Formal constructs and data visualization

To transform the rental data into meaningful input for the regression analysis and subsequently use the results for decision support, we first need to define the area of interest, i.e. the operating area of the carsharing provider within the city. For this purpose, we define a spatial area $A_i$ as a closed polygon of at least 3 different GPS-points, given by the tuple

$$A_i = (a_{i1}, \ldots, a_{in}, a_{i1})$$

Fig. 1. Schematic representation of methodology.

Fig. 2. Causal model relating POIs to carsharing activity.
\( |A_i| = n + 1 \geq 4 \),

\( \sigma'_i = (\phi'_i, \lambda'_i) \).

Eq. (2) states that the number of elements within such a tuple has to be at least four, while each element contains the latitude \( \phi \) and longitude \( \lambda \) describing a corner point of the polygon (Eq. 3). In principle, the operating area of an FFC provider, in which customers are allowed to end their rentals, is described by at least one such polygon. However, as illustrated in Fig. 3, which depicts the operating area of the carsharing partner in Berlin in April 2012, parts of the operating area can be disconnected. Hence, we define the operating area \( O \) as a set of \( m \) polygons \( O = \{A_1, A_2, \ldots, A_m\} \).

The blue polygons in Fig. 3 define the operating area at the start of the observation period. The red polygons show the expansion in November 2012, which we use later in the paper to validate our
approach. While customers are allowed to drive across the area borders and even park the car outside the areas during a trip, the rental can only be ended within the operation area. Thus, the operator can exert some degree of fleet control by excluding specific streets or even entire districts from the operating area. Nevertheless, the ability to guide customers with respect to where rentals end is quite limited in FFC systems, since this flexibility is their most appealing feature.

The data on rentals is also expressed using tuples, as

\[ r_q = (\varphi_{\text{start}}^q, \varphi_{\text{end}}^q, t_{\text{start}}^q, t_{\text{end}}^q). \]  

Here \( \varphi_{\text{start}}^q \) and \( \varphi_{\text{end}}^q \) represent the latitude and longitude of the staring location and destination, respectively, while \( t_{\text{start}}^q \) and \( t_{\text{end}}^q \) contain the timestamps of the start and end times of the rental.

Using the geographical information, we are subsequently able to visualize the driving behavior of carsharing customers. Fig. 4 illustrates a sample business day, during which typically around 3000 rentals are conducted. Within the illustration, a cross represents the starting point of a rental, while a square marks the destination. Naturally, the city center is the most vibrant region. However, the imbalances necessitate the use of relocation strategies following day or customers that continue travel by public transport.

In order to associate the number of rentals in particular locations with the surrounding POIs, we construct a grid \( G \) over the operation area and adjacent regions as follows:

\[ G(\Delta \varphi, \Delta \lambda, x, y) = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,y} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,y} \\ \vdots & \vdots & \ddots & \vdots \\ g_{x,1} & g_{x,2} & \cdots & g_{x,y} \end{bmatrix}, \quad x, y \in \mathbb{N} \]  

As Eq. (9) illustrates, the grid is defined as a matrix of \( x \times y \) tiles, with \( g_{ij} \) containing the geographical latitude and longitude of the tile center. The latitudinal and longitudinal edge lengths are given by \( \Delta \varphi \) and \( \Delta \lambda \), respectively. Hence, the latitude values of two longitudinally adjacent tiles are determined by the relationship \( \varphi_{i+1,j} = \varphi_{ij} + \Delta \varphi \). Given the spherical nature of the Earth, the resulting tiles are not perfect squares; however, this effect can be neglected for the comparatively small distances under consideration in this paper. The edge lengths consequently determine the granularity of the grid and of our analysis. A higher granularity (shorter edge length) is, however, not always better. While smaller edge lengths allow us to capture variations within a neighborhood, choosing edge lengths that are too small only results in a very large number of tiles whose features barely differ and an increased computational complexity.

Fig. 3 illustrates the temporal distribution of trips during the course of a day for each day of the week. It includes half a million trips undertaken within six months (before and after the expansion). As the figure shows, all days of the week exhibit very similar patterns. The only exception is a slight increase during the very early hours (midnight to 4 a.m.) of Saturdays and Sundays compared to the rest of the week. This is likely to be associated with social activities over the weekend. This difference between weekdays and the weekend must be considered when determining the operation area and implementing relocation strategies. For the proof of concept presented in this study, we focus on weekdays only. However, in a business setting the methodology can be applied to a weekday data set, as well as weekend data set, and managers can determine future expansions after weighing the results of both.

The third component of our analysis, besides operation area and rentals, are POIs. All POIs belong to the set \( P \) and are governed by

\[ P = \{ p_i \}_{1 \leq i \leq l}, \]  

\[ p_i \mapsto (\varphi_i, \lambda_i, \Gamma^i). \]  

A POI \( p_i \) is defined as a tuple consisting of its geographical latitude and longitude, as well as the set of categories it is tagged with \( \Gamma^i \). This set is a subset of the set of all categories \( \Gamma \) with 92 categories, i.e., \( |\Gamma| = 92 \).

\( \Gamma = \{ \text{accounting, airport, ..., zoo} \}, \)

\( \Gamma^i \subseteq \Gamma. \)

Fig. 5 illustrates the shares of rentals per time of day (total number of trips in parentheses).

**Fig. 5.** Shares of rentals per time of day (total number of trips in parentheses).
whether a specific location is inside $O$ in linear time. For each tile $g_{ij} \in G$ and each polygon $A_k \in O$ an arbitrary ray directly to the right from the center of $g_{ij}$ is drawn and the number of times this ray crosses the polygon $A_k$ of consecutive distinct vertices is counted. The algorithm returns true if and only if this number is odd, otherwise it returns false. If the test returns true we know that the tile lies in at least one of the areas $A_k \in O$. The resulting subset $G' \subset G$ of tiles includes all tiles that are also part of the operation area in Fig. 3. Overall, this reduces the number of tiles to be considered in the regression analysis to $|G'| = 24,280$.

Naturally, the total number of rentals in each tile constitutes the dependent variable in our regression. Accounting for the influence of points of interest is not as easily done, since a particular POI may influence the vehicle demand in several surrounding tiles – especially when considering the very small tile size of 100 by 100 meters. For instance, the parking spaces in front of a restaurant may be taken, so that the car is parked farther away, in an adjacent tile. The restaurant is, nevertheless, still the destination of the carsharing customers and the cause of vehicle demand on their return trip. Therefore, we do not only take the POIs within the tile into account, but also those within the vicinity of the tile. The model by Van der Goot [43] restricts such a vicinity by a maximum walking time of 40 min. Within this time span, the attractiveness of a POI is linearly decreasing, which represents the “willingness to walk” of the respective user and also reflects Tobler’s First Law of Geography – “Everything is related to everything else, but near things are more related than distant things” [47]. However, given the context of this paper, a 40 min upper bound is unrealistic. It would imply that parking 20 min from a shopping mall or restaurant is still attractive as parking directly in front the POI. Hence, we substantiate that the attractiveness of parking 20 min from a shopping mall or restaurant is still half as attractive as parking directly in front the POI.

As a final step, we sum up the relevance values for all POIs belonging to a specific category to receive the density of POIs belonging to category $k$ in tile $g_{ij}$. Hence,

$$\delta_{ijk} = \frac{1}{p_0 \cdot \delta_i \cdot g_{ij}} \sum_{p_{hi} \in \Gamma_i} \tau(p_{hi}, g_{ij}).$$

These density values, supplemented by demographic control variables, constitute the set of explanatory variables in our regression analysis. Essentially, instead of saying that there are 10 bars within 1000 m of the center of tile $g_{ij}$, we consider the distance between each bar and the tile. This is of great importance if, for instance, all bars are at a distance of 999 m. The resulting density $\delta_{ij,\text{bar}}$, referred to as the “bar-density at tile $g_{ij}$”, is only 0.016. Assuming now a different tile $g_{i,j'}$ that is only 10 m from all bars, it would have a $\delta_{i,j',\text{bar}}$ value of 9.998. Despite that both tiles enclose the same 10 bars within a range of 1 km, the bar-density of tile $g_{i,j}$ is more than 600 times higher relative to the other.

The control variables were only available on a neighborhood-level. Accordingly, the values of the control variables for a specific tile were derived from the neighborhood the tile is located in.

### 4.2. Regression analysis

The constructs defined in the previous subsection constitute the dependent and explanatory variables of our regression analysis. As the dependent variable $d_{ij}$, we use the number of rentals ending in a particular tile $d_{ij}$. To determine whether a rental $r_\gamma$ ends in a specific tile $g_{ij}$, we again employ the point in polygon algorithm. Applying this to all rentals, we receive for any tile the number of rentals that have ended in this tile. The resulting vector of dependent variables is $d = (d_{ij,1}, d_{ij,2}, \ldots, d_{ij,1}, d_{ij,2}, \ldots, d_{ij,3})^T$. The explanatory variables of our model are collected in the covariate matrix $C$ (Eq. 16). Besides a 1 for the intercept, this matrix contains all POI densities as well as all control variables. The numerical index of the densities represents the position of category $k$ in set $\Gamma$ given an alphabetical ordering. Hence, $\delta_{ij,1}$ is the density of POIs tagged with “accounting” in tile $g_{i,2,3}$, because “accounting” is the first element in an alphabetical ordering of all elements of set $\Gamma$. The variables $z_{x,y,1}$ to $z_{x,y,9}$ are control variables, such as population density, education, or income per person, provided by the city of Berlin. Thus,

$$C = \begin{bmatrix}
1 & \delta_{1,1,1} & \ldots & \delta_{1,1,1} & \tau_{1,1,1} & \ldots & \tau_{1,1,1} \\
1 & \delta_{1,2,1} & \ldots & \delta_{1,2,1} & \tau_{1,2,1} & \ldots & \tau_{1,2,1} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & \delta_{x,y,1} & \ldots & \delta_{x,y,1} & \tau_{x,y,1} & \ldots & \tau_{x,y,1}
\end{bmatrix}.$$
Consequently, the regression coefficients are given by the vector \( \mathbf{\beta} = (\beta_0, \beta_1, \ldots, \beta_{|\Gamma|+1}, \ldots, \beta_{|\Gamma|+m}) \), where \( \beta_0 \) is the coefficient for the intercept, \( \beta_1 \) to \( \beta_{|\Gamma|} \) for the POI categories, and \( \beta_{|\Gamma|+1} \) to \( \beta_{|\Gamma|+m} \) for the control variables.

Having constructed the sets of dependent and explanatory variables, we proceed with the regression analysis using a zero-inflated count model. The choice of the zero-inflated model becomes clear when considering the histogram of rentals over the grid, as illustrated in Fig. 7. The histogram shows the distribution of rentals per tile and should be read as e.g. “During the observation period (3 months, weekdays), 1332 tiles had three rentals ending there”. We can observe that most of the cells do not include any rentals, which is caused by various reasons. On the one hand, a particular tile might simply not be accessible by car, such as pedestrian zones, parks, or lakes. On the other hand, some locations, such as highways, may simply be generally uninteresting to end a rental at. Finally, the amount of zeroes is also caused by the high granularity of the constructed grid. Overall, the total share of tiles containing zeroes is approximately 42 percent. Together with the significance of the explanatory variables, the share of zeroes substantially exceeds the mean. Therefore, a standard count model using a Poisson distribution cannot be applied, since this distribution requires that the variance and mean are approximately equal.

A zero-inflated count model alleviates this issue by explaining the dependent variable through two separate processes that are simultaneously at work: a Poisson process and a zero-generating process responsible for generating excess zeroes [48]. Eqs. (17) and (18) illustrate this for a random variable \( X \), with \( \omega \) as the probability of a zero and \( s \in \mathbb{N}^+ \) as a positive integer.

\[
\Pr(X = 0) = \omega + (1 - \omega)e^{-\lambda}
\]

(17)

\[
\Pr(X = s) = \frac{(1 - \omega)e^{-\lambda} \lambda^s}{s!}
\]

(18)

When applying this regression model to our dataset, we have to be aware of the issue of multicollinearity. Naturally, some of the 92 POI categories frequently emerge together, such as different stores, shops, and malls, or even built-in combinations, like ATMs and banks. Therefore, multicollinearity between different categories inevitably exists. To alleviate this problem, we calculate the variance inflation factor (VIF) for each POI category. By using this criterion, we stepwise delete categories until none of the remaining variables exceed the generally used cut-off value of VIF = 5. Several general categories, like food, restaurant, or establishment were thus removed, since they are composed of and their influence is better explained by categories that are more specific. In total, 35 categories were deleted because of multicollinearity. Five additional categories were manually excluded, due to an insufficient number of observations. This includes categories such as “local government building”, for which we found very few POIs tagged with that specific category. Since we would expect a much higher number of such POIs in a city of more than three million inhabitants, only a small selection of these buildings is included in our data set. This could potentially bias our results and we consequently exclude categories, for which the expected number of POIs tagged with this category and the actual number substantially differ.

Employing the pscl package in R on our data, the zero-inflated model is fitted using maximum-likelihood. As the output, we receive two regressions. The first regression is the count model and returns the logarithm of the expected number of rentals for each tile given the covariates (Eq. 19). The second regression returns \( \pi_{ij} \), the logit of the probability \( a_{ij} \) that the number of rentals in a cell is zero and caused by the zero-generating process for each tile given the covariates (Eqs. 20 and 21).

\[
E(\ln d) = C \beta
\]

(19)

\[
E(\pi) = C \alpha
\]

(20)

\[
\pi_{ij} = \ln \frac{a_{ij}}{1 - a_{ij}}
\]

(21)

Table 1 displays the estimates, standard errors, z-values and significance levels of selected regression output coefficients (for the count regression). This selection includes control variables and significant POI categories.

### 4.3 Evaluation

Taking the results in Table 1 into account, we can make several observations regarding the first hypothesis outlined at the start of our paper. As before, the regression results refer to a period of three months with an overall amount of more than 150,000 rentals. In addition to several statistically significant control variables, many POI categories exhibit a strong significance, as well. This demonstrates that POIs bring an added value to the analysis by explaining effects that cannot be captured by the control variables. While the control variables outline the demographic composition of a certain area, POIs represent what draws people from the outside to this specific area. Hence, both variable types capture different effects, emphasized by both types having significant coefficients. Furthermore, the directions of the effects are theoretically sound. For instance, the coefficient for airport density is significantly positive and the airports are among the most active carsharing regions in Berlin. Furthermore, busses as means of short distance public transportation are significantly positive. These results confirm the findings of Katzev [49], who shows that carsharing is complemented by public transportation such as bus riding, as well as...
means of long distance and intercontinental travel. We also observe a significantly positive relationship of carsharing with entertainment facilities, such as movie theaters or night clubs, and places for recreation, such as spas. These are destinations where customers spend several hours. The rental is likely ended when exiting the vehicle, as opposed to destinations where customers spend only a few minutes, such as post offices or movie rental places. The coefficient of these POIs is consequently significantly negative, since the reservation is probably kept active to subsequently continue the trip.

With respect to control variables, a high population density, a low income, and a high share of foreigners all increase the attractiveness of the respective locations as carsharing destinations. This is reasonable, as the income of long distance and intercontinental travel. We also observe a

they are likely to stay in the city for on average shorter time periods than natives.

While these results indicate that some POIs do have a strong influence on carsharing activity, the overall explanatory power of the model, as captured in a pseudo-$R^2$ of 0.1, is relatively low. We investigate this discrepancy by conducting a visual analysis of expected and actual amount of rentals, as illustrated in Fig. 8. Both maps show the operation area, including areas for future extensions, as planned by the carsharing provider, (marked by red lines). Fig. 8b also depicts the number of actual rentals using a heat scale from transparent (zero rentals) to dark green (fifty or more rentals). Similarly, the values for the heat scale in Fig. 8a are constructed by multiplying the expected number of rentals from the Poisson-part of the regression by the probability that the tile contains a zero from the zero-generating process.

Fig. 8 simultaneously provides evidence that POIs correlate with the carsharing activity in Berlin, as well as an explanation for the low pseudo-$R^2$ of the ZINF-model. On the one hand, the left map (Fig. 8a) essentially mirrors its opposite, with the dark green regions of high activity being equally visible in both maps, just as well as unattractive areas. On the other hand, the strong contrast between each map is also visible. In Fig. 8a, transitions between areas of high and low activity occur over the span of several tiles (due to the relevance function in Eq. (11)), while such transitions are more sudden in Fig. 8b, with low-activity tiles occurring even in generally dark-green regions. The reason for this is simply that the prediction does not take traffic and accessibility issues into account – a generally attractive area might simply be inaccessible by car, such as pedestrian zones. It also explains the low explanatory power of the model, despite the very significant coefficients of some POI categories. However, it does not challenge the general influence of POIs on carsharing activity as proposed in Hypothesis 1. In fact, the visual similarities between predicted and actual regions of high, respectively low, activity as shown in Fig. 8 strongly support this hypothesis. In the next section, we will investigate whether this relationship also hold for areas that are not part of the initial expansion area at the left margin of either map in Fig. 8a.

### Table 1
Zero inflated Poisson regression (count model).

<table>
<thead>
<tr>
<th>Dependent variable: number of end rentals per tile, 24,280 observations (pseudo-$R^2$: 0.1)</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z-Value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.763</td>
<td>0.128</td>
<td>5.939</td>
<td>**</td>
</tr>
<tr>
<td>Airport</td>
<td>0.130</td>
<td>0.021</td>
<td>6.074</td>
<td>**</td>
</tr>
<tr>
<td>ATM</td>
<td>0.014</td>
<td>0.004</td>
<td>3.172</td>
<td>**</td>
</tr>
<tr>
<td>Bus station</td>
<td>0.024</td>
<td>0.002</td>
<td>12.827</td>
<td>**</td>
</tr>
<tr>
<td>Car rental</td>
<td>0.022</td>
<td>0.002</td>
<td>10.361</td>
<td>**</td>
</tr>
<tr>
<td>Car wash</td>
<td>-0.022</td>
<td>0.005</td>
<td>-4.601</td>
<td>**</td>
</tr>
<tr>
<td>Meal delivery</td>
<td>0.006</td>
<td>0.003</td>
<td>1.705</td>
<td>**</td>
</tr>
<tr>
<td>Meal takeaway</td>
<td>0.048</td>
<td>0.003</td>
<td>16.561</td>
<td>**</td>
</tr>
<tr>
<td>Movie rental</td>
<td>-0.034</td>
<td>0.007</td>
<td>-5.285</td>
<td>**</td>
</tr>
<tr>
<td>Movie theater</td>
<td>0.017</td>
<td>0.005</td>
<td>3.642</td>
<td>**</td>
</tr>
<tr>
<td>Night club</td>
<td>0.021</td>
<td>0.002</td>
<td>13.805</td>
<td>**</td>
</tr>
<tr>
<td>Parking</td>
<td>0.027</td>
<td>0.007</td>
<td>3.976</td>
<td>**</td>
</tr>
<tr>
<td>Post office</td>
<td>-0.041</td>
<td>0.008</td>
<td>-5.285</td>
<td>**</td>
</tr>
<tr>
<td>Shopping mall</td>
<td>0.030</td>
<td>0.006</td>
<td>5.483</td>
<td>**</td>
</tr>
<tr>
<td>Spa</td>
<td>0.036</td>
<td>0.004</td>
<td>8.718</td>
<td>**</td>
</tr>
<tr>
<td>Train station</td>
<td>-0.013</td>
<td>0.002</td>
<td>-5.194</td>
<td>**</td>
</tr>
<tr>
<td>Distance to city center</td>
<td>0.012</td>
<td>0.002</td>
<td>5.401</td>
<td>**</td>
</tr>
<tr>
<td>Share of foreigners</td>
<td>0.065</td>
<td>0.015</td>
<td>4.377</td>
<td>**</td>
</tr>
<tr>
<td>Age between 15–45</td>
<td>-0.194</td>
<td>0.152</td>
<td>-1.281</td>
<td>**</td>
</tr>
<tr>
<td>High education</td>
<td>-0.094</td>
<td>0.191</td>
<td>-0.987</td>
<td>**</td>
</tr>
<tr>
<td>Income &lt; 500</td>
<td>2.424</td>
<td>0.498</td>
<td>4.868</td>
<td>**</td>
</tr>
<tr>
<td>Log population density</td>
<td>0.260</td>
<td>0.036</td>
<td>7.351</td>
<td>**</td>
</tr>
</tbody>
</table>

Significance indicated at 15% (**), 10% (*), 1% (**), and 5% (***) levels.

**Fig. 8.** Predicted vs. actual amount of rentals (a) predicted and (b) actual. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)
magnified in Fig. 9. To avoid any issues arising from the initial period directly after the expansion, such as customers being unaware of it, we concentrate our analysis on the set of rentals between February and April 2013. Since the rental information of the new areas did not contribute to the regression from the previous section, we can use it now to assess whether the produced coefficients are robust when applied outside their original training data.

In Fig. 9, we compare predicted and actual number of rentals according to the same heat scale as in Fig. 8. There are several interesting observations, marked by ellipsoids and numbered. According to our predictions, the north-east corner of the original operation area (“1”) should exhibit the highest number of rentals, with the adjacent areas to the south-west and north of it (“2”) being slightly less active. These expectations are generally confirmed in Fig. 9b, with the area marked by “1” showing the darkest shades of green, associated with a high amount of rentals. Shades in the areas marked by “2” become lighter, but are still associated with more rentals than the other regions. However, deviations are also evident (e.g. “3”), all of which are occurring at the edges of the operation area. These are not predicted by our method, as they are not associated with POIs. Customers simply use the car to drive as far as allowed and consequently end the rentals at the border of the operation area. For strategic decision-making it may actually be a good thing that the impact of these rentals is filtered out by our approach, since the edges of the operation area will always be attractive, but do not represent the attractiveness of the underlying region. They may bias the results of the empirical analysis. However, they must be reconsidered later in the decision-making process.

Overall, this case study of a real-world expansion of the carsharing operation area demonstrates the value of our POI-centric approach in providing strategic decision support to carsharing providers, thereby supporting our second hypothesis. The Business Analytics tool, whose underlying method has been introduced in this paper, provides accurate predictions of expected carsharing activity in prospective locations. It condenses this information in a manner that can be easily interpreted and incorporated into managerial decisions – the heat maps illustrated in Figs. 8 and 9. If the business area shown in Fig. 9 had been constructed according to the predictions of our tool, it might have been limited to the areas marked by “1” and “2” – those areas that turned out to exhibit the by far highest level of carsharing activity in the region.

6. Conclusion

During the past decades, the popularity of carsharing as an alternative transportation service has continuously increased, turning it into one of the most promising business models for a more sustainable transportation sector. Increasing costs of fuel and maintenance, taxes, and environmental concerns are just a few of the reasons why people are switching from car ownership to carsharing. Companies in this quickly expanding market are constantly required to reassess business strategies, expand business areas, and react to shifts in customer demand.

In this paper, we develop a novel method to support this decision-making process. Our research is guided by two hypotheses. First, points of interest, such as restaurants, stores, or entertainment venues, can explain spatial variation in carsharing activity. Second, this relationship can be applied to assess prospective business area expansions. Building upon a large dataset provided by a globally leading carsharing service, we first investigate the influence of points of interest in an area on the attractiveness of that area for carsharing customers. We find substantial evidence for such a statistically significant relationship, supporting our first hypothesis. We investigate the second hypothesis using data from a real-world expansion of the original operation area. We show that our method is able to visualize the expected demand in the new area accurately when compared to the actual demand after the expansion. The methodology is implemented as a software prototype that can be applied to any city points of interest and carsharing data is accessible for. Our approach thus provides valuable information to carsharing enterprises of any size – from local startups to global players.

Our results also emphasize the potential of using information provided by services such as Google Maps and OpenStreetMap to explain spatial variation in human behavior. While carsharing is an interesting case study in an emerging business field, our approach is also applicable to other location-based services, such as advertisement and food-delivery. Furthermore, in our prior research, we
have used this methodology to investigate the relationship between criminal activity and geo-referenced social media activity [51] activities, and to determine the ideal spatial distribution of charge points for electric vehicles [52].

The objective of the software prototype constructed for this study is to provide decision support for the carsharing company; it does not automate decision-making. Carsharing is a complex business and humans can anticipate certain effects, such as the buzz created by the opening of a new shopping center, before they are reflected in the data. This complexity also provides several paths how future research can further augment our methodology. For instance, we will connect the insights from our study on charging infrastructure with the results of this paper on carsharing. In some cities, carsharing fleets are partially or completely electrified. We will investigate whether there is a significant difference in driving behavior between electric vehicles and those powered by combustion engines. Particularly the relevance of charge points as a new POI category will be taken into account.

Furthermore, the methodology introduced in this paper can be refined by analyzing how geographical features (traffic lights, stop signs, speed limits), traffic, and accessibility issues (parks or pedestrian zones) can be incorporated into the model to provide better estimates at a high resolution. The link between explaining demand and predicting demand in the model also needs further exploration [50]. Finally, as we have outlined in the beginning of this paper, the methodology can be adapted to provide insights for day-to-day operations. Hence, future research should analyze how the gained information can be used to derive optimal relocation strategies for carsharing vehicles to reduce idle time and save costs.

Acknowledgments

The authors would like to thank the guest editors, Michael Dourmpos and Constantin Zopounidis, as well as two anonymous reviewers for their outstanding work and valuable comments on our initial submission to the special issue. We also want to extend our gratitude to track chairs, reviewers, and discussants at the ICIS and HICSS conferences, where preliminary results of our study were presented, for valuable feedback. Tobias Brandt was supported in this research by a doctoral fellowship granted by the Foundation of German Business (sdw).

References


[31] Goh C, Millard-Ball A. Relocation simulation model for multiple-station carsharing systems. Transportation Research Record 2014;2416(-1):37