SCIENTIFIC REPORTS

OPEN

Quantifying Retail Agglomeration using Diverse Spatial Data

Duccio Piovani[1], Vassilis Zachariadis[2] & Michael Batty[1]

Received: 29 November 2016
Accepted: 26 May 2017
Published online: 14 July 2017

Newly available data on the spatial distribution of retail activities in cities makes it possible to build models formalized at the level of the single retailer. Current models tackle consumer location choices at an aggregate level and the opportunity new data offers for modeling at the retail unit level lacks an appropriate theoretical framework. The model we present here helps to address these issues. Based on random utility theory, we have built it around the idea of quantifying the role of floor-space and agglomeration in retail location choice. We test this model on the inner area of Greater London. The results are consistent with a super linear scaling of a retailer's attractiveness with its floorspace, and with an agglomeration effect approximated as the total retail floorspace within a 300 m radius from each shop. Our model illustrates many of the issues involved in testing and validating urban simulation models involving spatial data and its aggregation to different spatial scales.

Our approach to understanding and studying cities has radically changed in the last two decades[1,2]. Advances in spatial network theory[3,4], dramatic increases in the volume and type of available data[5] and insightful approaches based on out-of-equilibrium concepts[6,7] have contributed to the formation of a new, highly interdisciplinary *science of cities*. It is common knowledge that a fundamental role in the urbanization process is played by retail activity and indeed a city's social life and physical shape is greatly influenced by such activities. There is continual motion in cities as people move to work and play, in restaurants, bars, grocery shops, shopping malls and so on. Understanding and describing the mechanisms that govern these activities and the processes that lead to their agglomeration is not only intriguing, but crucial to an understanding of how a city works. For this reason modeling retail location has been a pillar of urban simulation for a long time[8]. Researchers have attempted different types of approaches to characterize this phenomenon, from entropy maximizing models[9,10], to random utility models[11–13] to agent based approaches[14,15]. However, retail location and consumer location choice theory continue to draw heavily from fundamental ideas[16,17] in central place theory[18,19] and bid-rent theory[20]. Since the late 1990s, advances in the new economic geography[21,22] have reinvigorated the field and reorganized it, by considering economies of scale, cross-dependencies in markets (e.g. between labour, retail and housing) and forms of imperfect competition between firms[23]. However, apart from notable exceptions[24–26], fresh approaches have been slow in informing state-of-the-art modelling and engaging with mainstream location choice modelling based on discrete choice and random utility theory, one of the cornerstones of consumer preference theory within economics developed over the last 40 years. Here we present such a model of consumer location choice based on random utility theory designed to capture both internal and external economies of scale at the individual retailer level. This means that we will consider single retailers, rather than retail centres which is usually the case when applying this type of approach, and describe how their economic properties depend both on personal characteristics (internal economies) and on the interaction with the other neighbouring retailers (external economies). In order to formalise both internal and external economies we need to quantify how the rent paid by a single retailer is affected by its floor-space (its surface are measured in m^2), and by its proximity to other retailers. Measuring this second effect will allow us to define an interaction range between individual retailers and understand the role played by retail agglomeration in consumer choice.

In the following discussion, we will examine the model's theoretical foundations, and propose an implementation strategy which takes advantage of unconventional data sources that have recently become available for calibration and validation. We then review the model's output predictions. Our study focuses on central London and the data gives us information on the spatial distribution of the population and work places (from the 2011 Population Census), on work to retail and home to retail trips (from the London Travel Demand Survey 2012)

[1]Centre for Advanced Spatial Analysis (CASA), University College London (UCL), 90 Tottenham Court Road, London, W1T 4TJ, UK. [2]Prospective Labs Ltd, IDEA London, 69 Wilson Street, Shoreditch, London, EC2A 2BB, UK. Duccio Piovani and Vassilis Zachariadis contributed equally to this work. Correspondence and requests for materials should be addressed to D.P. (email: d.piovani@ucl.ac.uk)

which is be used to calibrate the cost function of our model, and on the position, size and paid rent of single retailers (from the Valuation Office Agency 2010 data) which we use to quantify the effect of internal and external economies. Finally we use data from Foursquare, a website where users keep track and comment places they have been to, to validate the results. A richer description of the datasets we use is found below in the Methods section. As we will see, this work shows the central role retail agglomeration plays in consumer choice, and how this is created by local interactions at the *microscopic* level of the single retailer. In our model, consumers are evaluating the attractiveness of each shop by considering its size and the retail activity with a 325 *m* perimeter around it. This is in line with observations reported in literature on the average length of walking trips.

Results

The Model. Using as input the spatial distribution of population and individual retailer locations in London, our goal is to build a model of consumer choice capable of predicting the *success* of a retailer to a high level of accuracy. This will be measured by the expected number of people that will visit the retail store, or in other words, by the fraction of the distribution of trips that the model predicts will flow towards each single retailer. As implied in the introduction we have considered the retail activities found in the VOA dataset, and have treated them with no distinction of category. Under this simplifying assumption all retailers interact among each other in the same way (see supplementary material section 1, for an in depth analysis of the data and the retail categories we have included in our analysis). In order to build our model we only consider a retailer's location in the city, and its floor-space, i.e. the area of its surface or its size. We start by defining the utility of consumer *i* shopping for product *p* from retailer *r* as

$$u_{ir} = u_p - p_r - c(d_{ir}) + \omega_{ir} \tag{1}$$

where u_p is the utility that comes by acquiring the product, p_r the price at which the product is sold by retailer *r*, $c(d_{ir})$ a generalised cost of traveling between *r* and *i* and where ω_{ir} is a random element that reflects the personal tastes of the consumer for location and product variation. As one can see eq. 1 is a random utility function.

As one can see eq. (1) is a random utility function. The standard way of solving such an equation is by assuming the random element ω_{ir} to be i.i.d. (independently and identically distributed) Gumble distributed. The choice of the distribution is given because of its closed form, which allows us to calculate the probability of *i* choosing *r* over all other *r*′ possibilities. Details of the derivation of the equation are found in the supplementary material section 2. For now one only needs to understand that we define the probability of consumer i shopping in r as the probability of its utility being the greatest over all other alternatives $P(u_{ir} > u_{ir'})$ $\forall\ r' \neq r$, which has the form

$$p_{i\to r} = \frac{\exp(-\beta_i u_{ir})}{\sum_{r'}\exp(-\beta_i u_{ir'})} = \frac{\exp(-\beta_i(u_p - p_r - c(d_{ir})))}{\sum_{r'}\exp(-\beta_i(u_p - p_{r'} - c(d_{ir'})))} \tag{2}$$

where β_i is the inverse of the standard deviation of the distribution of ω_{ir}. If we now consider $\beta_i = \beta$ for every consumer, eq. (2) will depend only on the location of consumer *i* and retailer *r*.

Eq. (2) implies that each retailer *r* offers one, and only one, type of product *p*. Therefore, each consumer *i* associates only one random utility component ω_{ir} with each retailer, reflecting the fit between this sole product variation and the preferences of the consumer. This assumption is quite unrealistic; typically retailers will stock several types of products. It is reasonable to assume that larger shops will stock larger varieties. Therefore, variety can be expressed as a function of floor-space. Following the work of Daly and Zachary[27] and of Ben-Akiva and Lerman in ref. 28, and assuming that the price is constant for every retailer in the system, namely $p_r = p = \text{cost}$, the probability of consumer *i* shopping with retailer *r* is

$$p_{i\to r} = \frac{f_r^{\alpha}\exp(-\beta\, c(d_{ir}))}{\sum_{r'} f_{r'}^{\alpha}\exp(-\beta\, c(d_{ir'}))} \tag{3}$$

where f_r is the floor-space of retailer *r* and the exponent α quantifies the correlation between the stochastic components and the product variation. The model in eq. (3) is very similar to the multinomial logit model presented by McFadden[29, 30] and to the entropy maximising equation introduced by Wilson[9]. Macroscopically, this suggests that the utility that consumer *i* gets from buying from retailer *r* is either a sub-linear ($\alpha < 1$), linear ($\alpha = 1$), or super-linear ($\alpha > 1$) function of the floor-space of retail unit r. These internal economies (or dis-economies) of scale emerge from the application of a transparent utility-based approach and reflect perceived opportunities at the level of the individual retailer. Eq. (3) captures the potential impact of size at the individual shop level. As such it is sufficient in describing flow patterns and activity distribution associated with shop-size variations. But as we will see in the following section, it fails to account for the concentration of retail activity in clusters (markets/shopping centers). In other words, eq. (3) cannot explain agglomeration effects in the spatial distribution of retail activity.

In order to explicitly introduce agglomeration into the model, we start by defining the attractiveness of a retailer as

$$A_r = f_r^{\alpha} + \sum_{r':d_{rr'}} f_{r'}^{\alpha}\exp(-\gamma d_{rr'}) \tag{4}$$

where $d_{rr'}$ is the distance between retailers, and γ is parameter that weights the role of the retailer's neighborhood in the consumer's perception. The perceived utility associated with a given retailer is influenced by the utility associated with the neighboring retailers. By substituting eq. (4) into eq. (3) we finally get the form of the distribution to be used in our model as:

$$p_{i\to r} = \frac{A_r \exp(-\beta\, c(d_{ir}))}{\sum_{r'} A_{r'} \exp(-\beta\, c(d_{ir'}))} \tag{5}$$

In the location choice model presented in eq. (5), we can see how the parameter α controls the internal and γ the external economies of scale. Once again α tells us if the utility scales sub-, super- or linearly with the retailer's floor-space, while the γ values tune the *interaction range* between retailers: $\gamma \to \infty$ implies that the utility of visiting a retailer is only a function of its own floor-space f_r, while in the opposite case, when $\gamma \to 0$, the utility of a retailer is equally shaped by all shops in its vicinity. Note that for $\alpha = 1$ and $\gamma \to \infty$, we derive the multinomial logit model. The model in eq. (5) is a particular case of the *Cross-Nested* Logit model (CNL)[31, 32] in which retailer r represents a nest, and the extent of the overlap between nests is controlled by γ.

Now exploiting eq. (5), we can define the modeled turnover, i.e. the fraction of consumers that will shop in a given destination, for each retailer r as

$$Y_r = \sum_i n_i \cdot p_{i\to r} = \sum_i \left(\frac{A_r \mathrm{e}^{-\beta C(d_{ir})}}{\sum_{r'} A_{r'} \mathrm{e}^{-\beta C(d_{ir'})}} \right) \cdot n_i \tag{6}$$

where we have considered the population concentrated in centroids i,and where n_i indicates their respective populations. Therefore the term $n_i \cdot p_{i\to r}$ tells us the fraction of the population of centroid i that will travel to retailer r, and depends both on the attractiveness A_r of the retailer and on the distance between the two points. It will then be possible to compare the results yielded by eq. (6) with the actual rents paid by retailers. Quantifying the two parameters α and γ provides us with insight into the economical relationships between retailers which is the main objective of this paper. To do this we have to first calibrate the cost function which expresses how consumers perceive costs related to the distance between their origin and a shop.

Calibration of Cost Function. We define a cost function $C(d, \lambda)$ as a Box-Cox transformation of the distance, namely

$$c(d, \lambda) = \frac{d^{\lambda} - 1}{\lambda} \tag{7}$$

where d is the distance between the consumer and the retailer's location. The Box-Cox function adds an extra modeling layer that maps objective costs into perceived costs. The curve of this transformation ranges from linear ($\lambda = 1$) to logarithmic ($\lambda \to 0$). In the former case the marginal effect of cost travel d on trip distribution is $\exp(-\beta * d)$, while in the latter it becomes $d^{-\beta}$. To calibrate the two parameters β and λ we use the London Travel Demand Survey (LTDS) database which contains 5004 home to retail and 2242 work to retail trips (see Methods for more details). These numbers are not sufficient to calibrate the location choice model in full; we thus cannot use this dataset to give value to all the parameters used in model. In fact the LTDS supplementary report (TfL 2011) suggests that the sample size is sufficient only for an Inner/Outer London spatial classification. Therefore the survey is only used to calibrate distance profiles i.e. the distribution of distance traveled for shopping from home and work. As an input for the population distribution, we have used London Population Census data at the Lower Super Output Layer level (LSOA), where we have considered the population concentrated on the centroids. As we can see below in the Methods section, the database includes coordinates of the LSOA centroids, population and working population.

We start by arbitrarily assigning values to the α and γ parameters, which sets the values for the retailers' *attractiveness* in eq. (4). We chose $\alpha = \{0.8, 1.0, 1.8\}$ in order to cover the 3 different scenarios of sub-linear, linear and super-linear scaling of attractiveness with floorspace, and $\gamma = \{0.005, 0.05, 0.5\}$ which bearing in mind that we are working in meters, corresponds to a decay length of $\{200\,m, 20\,m, 2\,m\}$ respectively.

Now exploiting eq. (5) we can define the probability the model generates a trip of distance d as

$$p_m(d) = \frac{\sum_k n_k \sum_{r:d<d_{rk}<d+b_d} A_r \cdot e^{-\beta C(\lambda, d_{rk})}}{\sum_{k'} n_{k'} \sum_r A_r \cdot e^{-\beta C(\lambda, d_{rk'})}} \tag{8}$$

where the b_d in the numerator is a conveniently defined bin. In other words eq. (8) tells us the fraction of trips that are of length d. To fix the β and λ values, we tune them so to maximize the likelihood equation

$$\mathbb{L}(\beta, \lambda) = \sum_d p_e(d) \cdot \ln(p_m(d)) \tag{9}$$

where d is the length of the trips and $p_e(d)$ is its observed distribution in the LTDS. We repeat this process on both types of trips, in Fig. (1) and for every couple (α, γ) in the previous list we obtain a couple (β, λ) that maximizes the likelihood. The results yielded for β and λ are robust for the different α, and γ inputs, and the variations are of ± 0.1. As we can see from Fig. (1) there is a good agreement in both cases between the distributions observed in the data and those generated by the model with parameter values as $\beta_h = 0.25$, $\lambda_h = 0.3$, $\beta_w = 0.35$, $\lambda_w = 0.25$.

Correlation with Data. Considering the two different types of trips, the total modeled turnover predicted by the model is

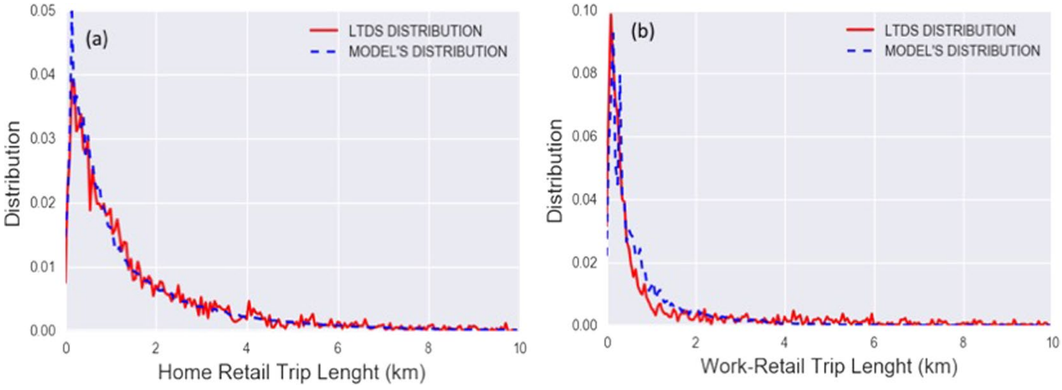

Figure 1. We show the maximum likelihood modeled distributions and the observed ones. For home-retail trips in (**a**) $\beta_h = 0.25$ and $\lambda_h = 0.3$, while for work-retail trips in (**b**) we have found $\beta_w = 0.35$ and $\lambda_w = 0.25$. As we can see from the figures in both cases the parameters that maximize the likelihood generate a distribution that fits the observed one very well.

$$Y_r = Y_r^w + Y_r^h = \sum_i (n_i^w p_{i\to r}^w + n_i^h p_{i\to r}^h) \tag{10}$$

and at an aggregated level this quantity can be used an indicator of the *success* of a retailer, with the idea that the more people visit a retailer the more that retailer earns. A more careful analysis should definitely take into account the type of goods sold by the retailer: indeed given the same number of visits, a car show room and a grocery corner shop yield different earnings, and a clothing shop in the vicinity of another clothing shop will have a different impact to that of a supermarket. These are two significant aspects that we leave for more detailed analysis in further research.

All the information on the retailers are found in the Valuation Office Agency database (see Methods VOA for more details), where we have found, among other data, information on the position, floorspace and the ratable value per meter for 98936 retailers in London. The rateable value represents the Valuation Office Agency's estimate of the open market annual rental value of a business/non-domestic property, i.e. the rent the property would let for on the valuation date, if it were being offered on the open market; as such, it is considered a very good indicator of the property value. Our hope is that the number of *visits* estimated by the model in eq. (10) to a retailer of any type, could be a good predictor of its rent. To test this hypothesis we calculate the Pearson correlation between the two quantities

$$Y_{\text{sq}}(\alpha, \gamma) = \frac{Y_r(\alpha, \gamma)}{f_r} \qquad R_{\text{sq}} = \frac{\text{Rateable Value}}{f_r} \tag{11}$$

where f_r is the floor-space. Not dividing by f_r would result in higher levels of correlation, not generated however through higher accuracy of the model but because of the variable f_r explicitly appearing in eq. (4), and because of the strong correlation between the two quantities

We therefore study the behavior of

$$C(\alpha, \gamma) = \frac{\text{cov}(Y_{\text{sq}}, R_{\text{sq}})}{\sigma_{Y_{\text{sq}}} \sigma_{R_{\text{sq}}}} \tag{12}$$

looking for the values $(\alpha_{\max}, \gamma_{\max})$ that maximize the correlation. These parameters will be considered those that best describe the retail activity through internal and external economies. In Fig. (2a)a we show a heat map of the correlations found exploring the parameter space. The maximum correlation is $C_{\max} = 0.48$ which seems like a reasonable value given the level of aggregation we are working at with the parameter values $(\alpha_{\max}, \gamma_{\max}) = (1.3, 0.008)$. We should stress at this point that we do not consider the correlation level obtained as a validation of the model. These values are those normally obtained by models of this type and probably depend more on density effects[33, 34] than on the details of our model. As we have said, our main goal is to quantify and weight of internal and external economies, and we will now do this characterising the values of the parameters that maximise the correlation. As we can see the $\alpha_{\max}$ tells us that a model with a super linear scaling between the number of visitors and floor-space best describes the data, while the $\gamma_{\max}$ value indicates that retailers benefit from their proximity to other retailers given a characteristic length of the decay of $\frac{1}{\gamma} = 125\,m$.

We can see in Fig. (2a) how for increasing values of γ, which imply a faster decay and therefore a shorter *interaction range* between retailers, the correlation levels decrease. In the extreme case of $\gamma = \infty$, the attractiveness term in eq. (4) becomes $A_r = f_r^\alpha$ meaning that retailer attractiveness only depends on its floor-space. An A_r of this form is what is usually found in standard retail models such as those in ref. 9, and works well when considering retail centers. From these results it seems that for a microscopical formalization of the retail activity, this

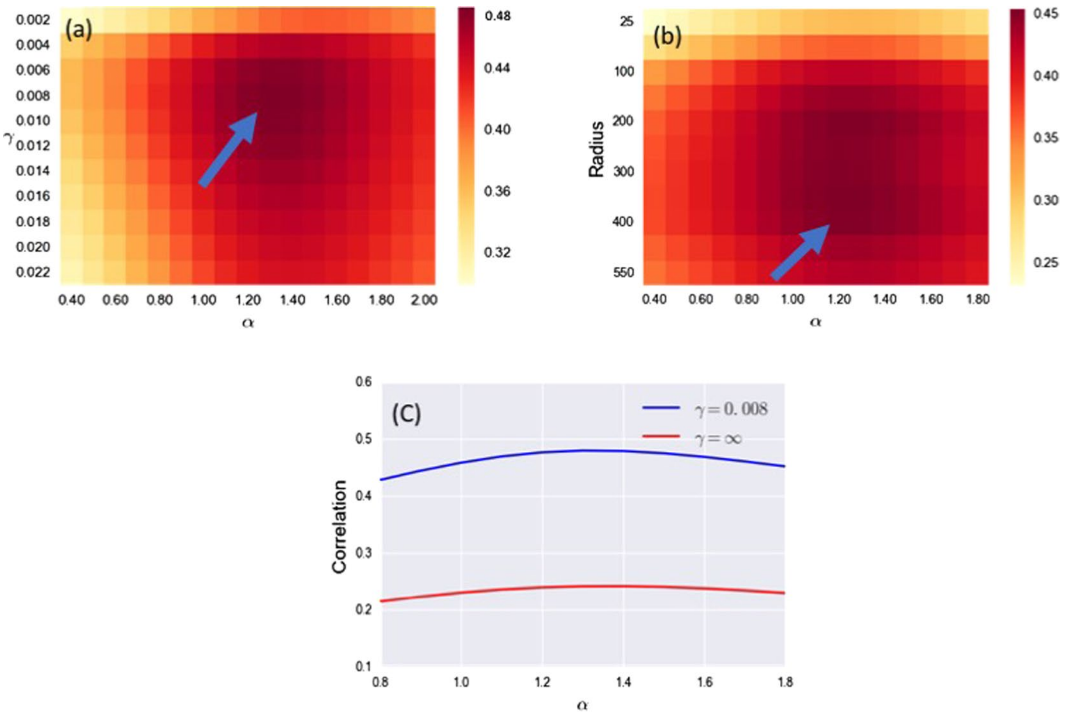

Figure 2. In this figure we present the results obtained comparing the retailer's ratable value found in the VOA dataset, with the modeled turnovers $Y(\alpha, \gamma)$. In (**a**) we show a heat map of the correlation between the two quantities, we can see a clear maximum in the correlation for $\alpha_{\max} = 1.3$ and $\gamma_{\max} = 0.008$. In (**b**) we show the correlation with the model of $\gamma = 0$, which implies that the attractiveness becomes $A_r = f_r^{\alpha} + \sum_{r':d_{rr'}<d_o} f_r^{\alpha}$, and we have studied the correlation for different values of d_o. We can see how the maximum is found in $\alpha_{\max} = 1.2$ and $d_{\max} = 325\,m$ which is 2.6 times the characteristic length of the decay and coincides with a dampening effect of 93% of the interactions. In (**c**) we compare the correlation in the case where $\gamma = \infty$, or with *no interactions*, which is the form used in other more aggregated models. We can see how in this microscopic formalization the model always correlates better with the data for finite values of γ.

picture fails as a satisfactory description. In Fig. (2c) we compare the correlations obtained from the model with *interactions* and with *no interactions*. We can see how both models exhibit the maximum for the same value of α and how the model with interaction has constantly higher correlation values. This means that agglomeration has to be included when modeling consumers' choice at the level of the single retailer.

Another interesting scenario is the opposite case with $\gamma = 0$, where consumers perceive the neighborhood of a retailer in the same way for a given distance. The attractiveness term in eq. (4) becomes

$$A_r = f_r^{\alpha} + \sum_{r':d_{rr'}<d_o} f_r^{\alpha} \tag{13}$$

where we have introduced a parameter d_o which explicitly sets the *interaction range* between retailers. We can now study the Pearson correlation for different values of the radius d_o and α. In Fig. (2b) we can see how the maximum values are found for $\alpha = 1.2$ which is very similar to what has been previously found for $d_o^{\max} = 325$, but despite the correlation levels being very close $C_{\max} = 0.45$ for the model with the exponential decay, this model typically shows lower values. What this seems to imply is that consumers implicitly take into account the distance and weight closer shops more than more distant ones, instead of equally considering their proximity.

In order to better quantify the level of agglomeration predicted by the model we can now compare the results in Fig. (2a,b). As implied, an exponent of $\gamma = 0.008$ indicates a characteristic decay length of $\frac{1}{\gamma} = 125\,m$ for the interaction range. If we compare this result with the $\gamma = 0$ model, we see that $d_o^{\max} = 2.6 \cdot \frac{1}{\gamma}$, and if we substitute it in the exponential, we get $e^{-0.008 d_o^{\max}} = 0.07$ which means that at that distance retailers loose 93% of the benefits of neighboring floor-space. This analysis suggests that retail activity benefits from agglomeration, and this benefit is spread through local interactions between single retailers, whose main effect is felt within 325 *m*. From the consumer's point of view, we could say that when choosing the retailer to visit, consumers take into account the amount of floor-space or choice within a radius of around 325 *m*. This result is in very good agreement with several studies on walking trips as in ref. 35 where it is shown that 65% of all walking trips are under 0.25 miles $\simeq$ 400 m, and in ref. 36 where the average walking distance for several types of retail activity are just above $\simeq$500 m. We interpret these results as follows: the choice perceived when considering a specific retailer is quantifiable as its floor-space, and the floor-space of retailers which can be reached by foot.

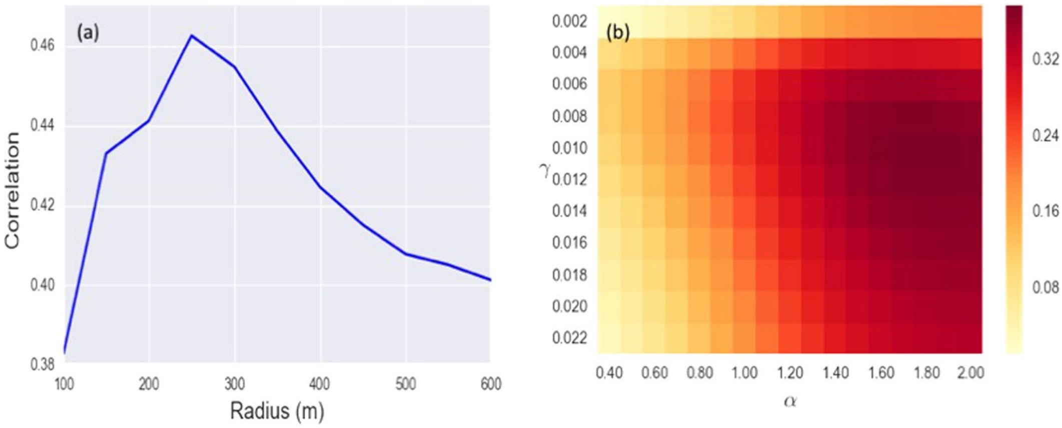

Figure 3. In the left panel in (**a**), we show the correlation between the rateable value of retailers in the VOA dataset, and the check-ins found in Foursquare around them for different radii. We can see how the correlation peaks at $R=250\,m$. In the right panel in (**b**) we show the correlation between the expected turnovers $Y(\alpha, \gamma)$ of each shop and the number of check-ins measured within a radius of $R=250\,m$ from it. The set of parameters that maximize the correlation are $\alpha_{\max}, \gamma_{\max}=(1.8, 0.01)$, which are quite similar to what we found for the VOA rateable values.

Discussion

In this section we will work through the tests we have done in order to validate the levels of agglomeration and the scaling law predicted by the model. We have done so by repeating the same procedure both on a randomized data set and on data coming from Foursquare, a web site where people check in, and comment bars restaurants and other retail activities (see supplementary material for more information). We have randomized the data both in the rents/m^2, and in their locational positions. To randomize the rents/m^2, we have first measured their distribution in the VOA dataset and then assigned a random rent to each retailer taken from the same measured distribution, while to randomize the position we have simply reshuffled the position of retailers taking as lower and upper bounds the minimum and maximum coordinates found in the data. In both cases the correlation between the rents and the *turnovers* predicted by the model disappeared completely ($C_{\max} \simeq 10^{-3}$). This seems to imply that there is a clear relation between the retailers' spatial interactions and the rents they pay, and that these are captured, at an aggregated level by eq. (4).

A further validation consists in testing the model on *check-in* data coming from Foursquare. As we can see in the Methods section, this data set gives us information on the number of visitors who checked in a given venue on the application. The dataset contains all Foursquare venues within the M25 motorway, which consists of 300×10^3 venues, of which we have filtered just above 100×10^3 retail activity (more information on this process is in the Methods section). Each record contains location coordinates, number of check-ins since the venue was registered together with a detailed venue category. For each retailer r in the VOA dataset, we have summed the number of check-ins that were registered in the Foursquare dataset in a range R from the retailer. To choose the range R, we have studied the correlation between the Foursquare check-ins and the rent/m^2 seen in the VOA dataset.

As we can see from Fig. (3a) the maximum level of correlation is obtained by fixing $R=250\,m$ as the range to count check-ins. Curiously enough this is quite close to the agglomeration characteristic exponential decay length. In Fig. (3b) we show the correlation between the modeled turnovers and the check-ins happening in range $R=250\,m$ from the retailers. We can see how the two parameters for which the correlation is maximum are ($\alpha_{\max}$, $\gamma_{\max}$) $=(1.8, 0.01)$ which, considering the big difference in the data used, can be considered incredibly similar to what we previously obtained. The α value is still in line with a super-linear scaling between floor-space and attractiveness, and the γ implies a characteristic decay length of $\frac{1}{\gamma} = 100.0\,m$, against the $\frac{1}{\gamma} = 125\,m$ previously predicted. Once again given the great difference between the two datasets we have used, we consider that these results are a reassuring sign of the robustness of our findings.

In order to understand the strength and weaknesses of our model, for each retailer we have studied and compared the difference between the modeled turnover $Y(\alpha_{\max}, \gamma_{\max})$ and the actual VOA rent. This will allow us to find out for which type of retailers the model is more accurate and which will have to be the future steps to better shape it. The first forced step is to make the two quantities comparable. We call $\mathbf{R}$ the vector of the *fraction* of rents, where the *ith* component $\frac{R_i}{\sum_i R_i}$ represents the fraction of the total rent in the system that belongs to the *ith* retailer. In the same way, we define the vector of the fraction of modeled turnovers $\mathbf{Y}$, with components $\frac{Y_i}{\sum_i Y_i}$. We can now define the E error vector as

$$\mathbf{E} = \log\left(\frac{\mathbf{Y}}{\mathbf{R}}\right) \tag{14}$$

The components of the *error vector* in eq. (14) will indicate an overestimation of the *success* of a retailer, if $E_i > 0$ and an underestimation in the opposite case $E_i < 0$. By observing the characteristics of the retailers for which we

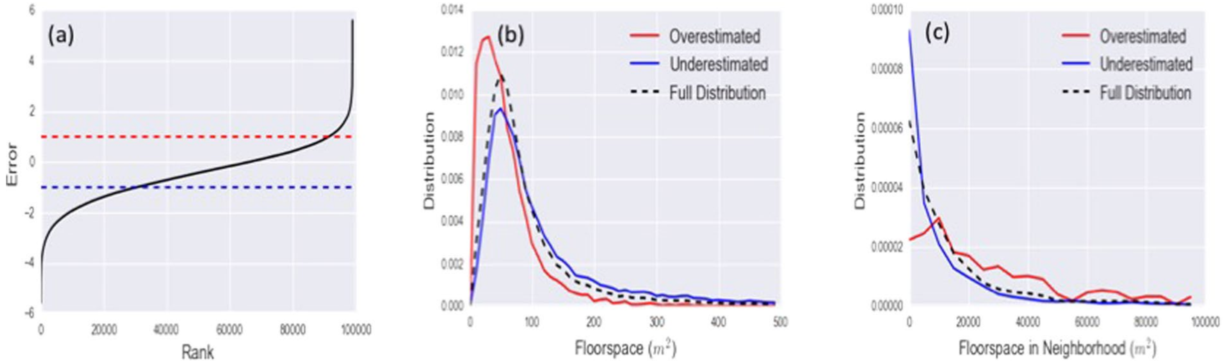

Figure 4. In this figure we present the results obtained analyzing the model's errors in estimating the retailers' *success*. In (**a**) we have ordered the errors by rank. The red and blue dashed lines indicate the thresholds we have used to define the subset of *overestimated* and *underestimated* retailers. In (**b**,**c**) we show that the overestimated retailers tend to be smaller than average with more floor-space than average in their neighborhood.

underestimate and overestimate the rent, we can see how patterns emerge. In Fig. (4a) we show a ranked distribution of the errors E_i. From the asymmetry of the curve we realize how the model, in this interpretation, tends to underestimate a retailer's success. Indeed for the vast majority of the retailers we analysed, we noted that the Y_i is smaller than their R_i which implies that the overestimation errors are typically larger. We can therefore study the characteristics of the retailers we have overestimated, those above the red dashed line, and compare them with those we have underestimated, below the blue line. Observing the red curves in Fig. (4b,c), we can see how the model tends to overestimate retailers with a *small* floor-space surrounded by many neighbors, which looking at our definition of attractiveness in eq. (4) is predictable. On the other hand we can see how the blue curves, which represent the distributions of the underestimated retailers are very similar to the the distribution of the full dataset (indicated by the black dashed line). From the figures it is clear that the model underestimates the number of visits to big retailers in locations with low retail activity concentration, such as supermarkets and other retailers that likely fall under convenience rather than comparison retail, where intra-store variety is more important than inter-store variety.

Outlook.

In this paper we have presented a location choice model, based on random utility theory and following the growing popularity of the cross-nested choice structure, we have tested this on unconventional data sources of different types. The novelty of the proposed model is that it simulates retailers at the individual level. This opens up exciting opportunities towards integrating the consumer location choice component with explicit retail location micro simulation models able to take full advantage of the emerging availability of detailed data sources while also incorporating complex behavior on price setting, network dynamics and risk management. The proposed model in its current form has been simplified (with no loss of generality) into assuming that all retailers offer unique varieties of the same product. Moreover, it has been assumed that (i) all consumers have equal disposable retail budgets regardless of their location, (ii) all trips are uni-purpose (only shopping is considered), (iii) VOA rateable values are good indicators of floor-space rents and (iv) product prices do not vary in space. These assumptions mean that a considerable part of the complexity of the decision making mechanism is not represented by the model. The VOA and LTDS datasets that we are currently using have the potential to increase the complexity of the model significantly towards removing some of the existing simplifying assumptions, and when combined with passively collected social media datasets and formal datasets on economic activity (e.g. from the Business Structure Dataset developed by ONS), these models could offer sufficient detail to capture all the main dimensions of behavioral variation. Having said this, the basic model that we present in this paper remains very useful both as the baseline example of the proposed approach and as a benchmark; despite its simplicity, it is able to capture and reproduce both internal and external economies of scale very closely. Further immediate lines of research could consider different categories of retailers separately as well as applying the same analysis on data coming from other cities. Looking at the not too distant future, passively generated datasets of human presence promise to offer deeper insights into the dynamics of urban activities, including spatio-temporal patterns of shopping behavior. Having said that, the abundance of existing social media datasources and the relentless pace in which new data are currently being introduced, sustain the promise of accessible and highly dis-aggregated spatiotemporal information for anyone who manages to overcome lack of specification, representational biases and possibly absence of context[37].

References

1. Batty, M. New ways of looking at cities. *Nature* **377**, 574 (1995).
2. Bettencourt, L. & West, G. A unified theory of urban living. *Nature* **467**, 912–913 (2010).
3. Barthélemy, M. Spatial networks. *Physics Reports* **499**, 1–101 (2011).
4. Crucitti, P., Latora, V. & Porta, S. Centrality measures in spatial networks of urban streets. *Physical Review E* **73**, 036125 (2006).
5. Batty, M. Big data, smart cities and city planning. *Dialogues in Human Geography* **3**, 274–279 (2013).
6. Louf, R. & Barthelemy, M. Modeling the polycentric transition of cities. *Physical Review Letters* **111**, 198702 (2013).
7. Louf, R. & Barthelemy, M. How congestion shapes cities: from mobility patterns to scaling. *Scientific Reports* **4** (2014).
8. Huff, D. L. A programmed solution for approximating an optimum retail location. *Land Economics* **42**, 293–303 (1966).

9. Wilson, A. G. Entropy in urban and regional modelling (2011).
10. Wilson, A. G. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy* **3**, 108–126 (1969).
11. Teller, C. & Reutterer, T. The evolving concept of retail attractiveness: What makes retail agglomerations attractive when customers shop at them? *Journal of Retailing and Consumer Services* **15**, 127–143 (2008).
12. Manchanda, P., Ansari, A. & Gupta, S. The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Science* **18**, 95–114 (1999).
13. Williams, H. C. On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A* **9**, 285–344 (1977).
14. Vanhaverbeke, L. & Macharis, C. An agent-based model of consumer mobility in a retail environment. *Procedia-Social and Behavioral Sciences* **20**, 186–196 (2011).
15. Heppenstall, A. J., Harland, K., Ross, A. N. & Olner, D. Simulating spatial dynamics and processes in a retail gasoline market: An agent-based modeling approach. *Transactions in GIS* **17**, 661–682 (2013).
16. von Thünen, J. H. & Hall, P. G. *The Isolated State, an English edition of Der isolierte Staat* (Pergamon, 1966).
17. Hotelling, H. Stability in competition. In *The Collected Economics Articles of Harold Hotelling*, 50–63 (Springer, 1990).
18. Christaller, W. *Central Places in Southern Germany* (Prentice-Hall, 1966).
19. Dennis, C., Marsland, D. & Cockett, T. Central place practice: shopping centre attractiveness measures, hinterland boundaries and the uk retail hierarchy. *Journal of Retailing and Consumer Services* **9**, 185–199 (2002).
20. Alonso, W. A theory of the urban land market. *Papers in Regional Science* **6**, 149–157 (1960).
21. Krugman, P. Increasing returns and economic geography. Tech. Rep., National Bureau of Economic Research (1990).
22. Krugman, P. What's new about the new economic geography? *Oxford Review of Economic Policy* **14**, 7–17 (1998).
23. Dixit, A. K. & Stiglitz, J. E. Monopolistic competition and optimum product diversity. *The American Economic Review* **67**, 297–308 (1977).
24. Anderson, S. P., De Palma, A. & Thisse, J. F. *Discrete Choice Theory of Product Differentiation* (MIT Press, 1992).
25. Suarez, A., del Bosque, I. R., Rodriguez-Poo, J. M. & Moral, I. Accounting for heterogeneity in shopping centre choice models. *Journal of Retailing and Consumer Services* **11**, 119–129 (2004).
26. Fiasconaro, A., Strano, E., Nicosia, V., Porta, S. & Latora, V. Spatio-temporal analysis of micro economic activities in rome reveals patterns of mixed-use urban evolution. *PloS One* **11**, e0151681 (2016).
27. Daly, A. & Zachary, S. Improved multiple choice models. *Determinants of Travel Choice* **335**, 357 (1978).
28. Ben-Akiva, M. E. & Lerman, S. R. *Discrete Choice Analysis*: *Theory and Application to Travel Demand*, vol. 9 (MIT Press, 1985).
29. McFadden, D. Economic choices. *The American Economic Review* **91**, 351–378 (2001).
30. McFadden, D. Econometric models for probabilistic choice among products. *Journal of Business* S13–S29 (1980).
31. Wen, C.-H. & Koppelman, F. S. The generalized nested logit model. *Transportation Research Part B: Methodological* **35**, 627–641 (2001).
32. Bierlaire, M. A theoretical analysis of the cross-nested logit model. *Annals of Operations Research* **144**, 287–300 (2006).
33. Um, J., Son, S.-W., Lee, S.-I., Jeong, H. & Kim, B. J. Scaling laws between population and facility densities. *Proceedings of the National Academy of Sciences* **106**, 14236–14240 (2009).
34. Pan, W., Ghoshal, G., Krumme, C., Cebrian, M. & Pentland, A. Urban characteristics attributable to density-driven tie formation. *Nature communications* **4** (2013).
35. Yang, Y. & Diez-Roux, A. V. Walking distance by trip purpose and population subgroups. *American journal of Preventive Medicine* **43**, 11–19 (2012).
36. Millward, H., Spinney, J. & Scott, D. Active-transport walking behavior: destinations, durations, distances. *Journal of Transport Geography* **28**, 101–110 (2013).
37. Zachariadis, V., Vargas-ruiz, C., Serras, J. & Ferguson, P. Decoding Retail Location: A Primer for the Age of Big Data and Social Media. *Cupum*, *2015* (2015).

Acknowledgements

All authors thank the EU FP7 for their support through the INSIGHT Project (Innovative Policy Modelling and Governance Tools for Sustainable Post-Crisis Urban Development) (http://www.insight-fp7.eu). M.B. wishes to thank the ERC for support through grant no. 249393-ERC-2009-AdG. D.P. acknowledges the support of the EPSRC grant RCUK EP/M023583/1 as part of the Digital Economy network and is grateful to C. Molinero and R. Morphet for continuous and useful discussions, and to A. Ialongo for a long and fruitful conversation on the analysis of the results. Authors would also like to acknowledge the contribution of C. Vargas-Ruiz, J. Serras and P. Ferguson in the development of the the model as described in ref. 37.

Author Contributions

D.P. and V.Z. contributed equally to the scientific research in this work. V.Z. developed the model, D.P. implemented it and performed the simulations, and they both analysed the results. M.B. directed the project as P.I. All authors were involved in the writing of the paper.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-05304-1

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.