Decoding Retail Location: A Primer for the Age of Big Data and Social Media

Vassilis Zachariadis, Camilo Vargas-Ruiz, Joan Serras, Peter Ferguson and Michael Batty

Abstract

The rise of social media and the open government movement are revolutionising data-driven research in social sciences. In this paper we use passively collected datasets of human mobility and public records of economic activity to develop a shopping location choice model that considers internal and external economies of scale at the outlet level. Our findings highlight the impact of scale and agglomeration on shopping location choices and form a platform for further research.

V. Zachariadis (Corresponding author) • C. Vargas-Ruiz, J. Serras, P. Ferguson, M. Batty
Centre for Advanced Spatial Analysis (CASA), University College London (UCL), London W1N 6TR, UK

Email: zachariadis@gmail.com
C. Vargas-Ruiz

Email: camilo.ruiz@ucl.ac.uk
J. Serras

Email: j.serras@ucl.ac.uk
P. Ferguson

Email: peter.ferguson.10@ucl.ac.uk
M. Batty
Email: m.batty@ucl.ac.uk

1 Introduction

From Huff (1966) to Teller and Reutterer (2008), modelling retail location choice processes has been one of the pillars of urban modelling. Recent work (Leszczyc et al, 2004; Reimers and Clulow, 2004) has shed empirical light to pricing strategy and location preference of retail activity. However, retail location and consumer location choice theory continue to draw heavily from fundamental ideas from von Thunen (1966) and Hotelling (1931), from central place theory (Christaller, 1966; Dennis et al, 2002) and rent-bid theory (Alonso 1960).

Admittedly, since the late 90s, advances in economic geography (Krugman, 1990 and 1998) have reinvigorated the field and reorganised it, by considering economies of scale, cross-dependencies of markets (e.g. labour, retail, housing) and forms of imperfect competition between firms (Dixit and Stiglitz, 1977). However, apart from notable exceptions (for example, Anderson et al, 1992; Suarez et al, 2004), 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 (Williams, 1977).

Here we present a model of consumer location choice, based on random utility theory and designed to capture internal and external economies of scale at the individual retailer level. We discuss its theoretical foundations, propose an implementation strategy, take advantage of unconventional data sources that became recently available (e.g. detailed economic activity, digital social media footprints etc.) for calibration and validation and review its outputs.

2 Datasets

Before we go on to present the proposed consumer location choice model, in this section we present the main datasets used for its calibration and validation. These are a combination of formal proprietary datasets of travel behaviour and economic activity, and passively collected data sources of digital social media footprints.

2.1 London Travel Damand Survey (LTDS)

LTDS is a continuous household survey of the London area, covering the London boroughs as well as the area outside Greater London but within the M25 motorway. Results in this report relate to residents of the Greater London area, comprising the 32 London boroughs and the City of London. The first year of results covered the financial year 2005/06, meaning that there are now eight years of data available.

The survey is a successor to the household survey component of the London Area Transport Survey (LATS) which was last carried out in 2001. The LTDS annual sample size is around 8,000 households in a typical year, a sum of 65,000 households for the 2005-2013 period.

LTDS captures information on households, people, trips and vehicles. All members of the household are surveyed, with complete trip detail for a single day recorded for all household members aged 5 and over. Three questionnaires are used a household questionnaire, individual questionnaires for all household members, and trip sheets or travel diaries. The later capture data on all trips made on a designated travel day, the same day for all members of the household. Details captured include trip purposes, modes used, trip start and end times, and the locations of trip origins and destinations.

2.2 Valuation Office Business Rates

There is large number of studies providing evidence for high correlation between flow volumes of shopping pedestrians and turnover of neighbouring retail outlets (Chiradia et al 2009, Timmermans and Waerden 1992, Thomas and Bromley 2003, Thornton et al 1991 etc.). As a result, the literature suggests that, in the absence of data, pedestrian flows are frequently used to estimate retail turnover and vice versa (Borgers et al 2008, Winrich 2008).

Since, the recent online publication of the Valuation Office Agency Business Rates for 2005 and 2010, the business rates of all business premises in England and Wales has become available to the public and offer a unique in depth, extent and geographic precision dataset. VOA compiles and maintains lists of rateable values of the 1.7 million non-domestic properties in England, and the 100,000 in Wales, to support the collection of around 25 billion in business rates.

The Rateable value represents the agencys 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.

The rateable value is estimated based on the varying rents at the vicinity of a property and decided to represent a reasonable level of open market rental value, taking into account the expected turnover of the premise, the size, age and condition of the property, the length of the frontage, the depth and vertical layout, the visibility, the footfall and pedestrian flow volumes on surrounding streets etc. Because the rateable value of a property is used to determine the non-domestic property tax (business rate), and the evaluators have access to detailed information (such as contracts, revenue documents etc.), and comprehensive evaluation documentation and methodology, rateable value is considered a very good indicator of the property value of a hereditament.

The agency publishes a detailed set of information for each property; this includes the classification of its main use (detailed breakdown into more than 100 classes), the full address and postcode, the total area of the premise, the total rateable value and breakdown into zones with different rateable value per square metre, and the weighted average rateable value per square metre. This makes it possible to create a detailed map of rateable value for any use.

2.3 Social media spatiotemporal profiles (Twitter and Foursquare)

We use two passively generated datasets. One contains 25 millions geo-located tweets collected over a period of 7 months (10/2013 to 05/2014), and covering an area that extends beyond London and covers most of the Greater South East of England. Each record contains the tweet coordinates, timestamp, text, language, tags etc. For this paper we use only location and time.

A second one contains all Foursquare venues within the M25 motorway (300 thousands venues). Each record contains location coordinates, number of check-ins and unique visitors since venue was registered and detailed venue category (activity type). The Foursquare venue data was collected in December 2014.

3 Methodology

In this section we present the shopping location choice model and the steps of the calibration process.

3.1 Shopping location model

The objective is to compose a causal model of actors (consumers and retailers) that explains the spatial distribution of retail activity. We assume that consumers choose shopping locations that maximise their utility and that retailers are willing to pay floorspace rent in line with the consumption that they expect to attract (eq. 1).

$$c_{r,j} = a_r \times x_{r,j} + b_r \tag{1}$$

where $c_{r,j}$ is the floorspace rent $(£/m^2)$ that retailer r is willing to pay in location j, $x_{r,j}$ is expected consumption per m^2 of floorspace and a_r, b_r are constants particular to retailer r. In the case of the baseline model of this paper, we assume b_r is zero for all retailers. This means that retailer r is willing to pay *ad valorem* rents a_r, which reflect gross profit margins etc. To simplify, for this paper we also set $a_r = a$ fixed for all retailers.

We assume that the net utility of consumer q in location i buying product type f from retailer r in j is equal to:

$$u_{f,q,i,r,j} = u_f - p_{f,r} - c_{f,i,j} \tag{2}$$

where u_f is the utility of acquiring f, $p_{f,r,j}$ is the price retailer r in j is selling f for, and $c_{f,i,j}$ is the cost of transporting f from i to j. For simplification we assume a sole homogeneous product type f. Therefore, the transportation costs $c_{f,i,j}$ per unit of product are equal for all varieties of product f; i.e. equal regardless of the retailer of choice.

Despite the homogeneity assumption, we follow Fujita et al (2001, see also Fujita and Thisse, 2013) in assuming monopolistic competition (Dixit and Stiglitz, 1977) between retailers. This means that each retailer offers a unique variety of products; i.e. each retail unit brings to the market a variety of products sufficiently differentiated to represent a unique blend of product type f. As a result, retailers hold the power to set their selling prices. Moreover, the number of retailers is assumed to be sufficiently high and the entry/exit costs to the market comparatively low, and therefore, profits to be zero and strategic decision-making non-existent.

Taking into account the spatial dimension of the problem, the source of product variation between retailers is a combination of differences in the actual product in offer and in the location of sale point (the shop) (Greenhut et al, 1987). Therefore, from eq. 2 we have:

$$u_{f,q,i,r,j} = (u_f + w_{q,f,r}) - p_{f,r} - c_{f,i,j} \tag{3}$$

where $w_{q,f,r}$ is a random element of utility, reflecting fit between the variation of f offered by retailer r and the particular preferences of consumer q in terms of spatial location and product variety. If random utility $w_{q,f,r}$ is Gumbel i.i.d. for all retailers then following Train (2003) the probability of consumer q in i choosing to shop from retailer r in j is:

$$P_{f,q,i,r,j} = \frac{\exp(\beta_q \times (u_f - p_{f,r} - c_{f,i,j}))}{\sum_{[r',j'] \in [R,J]} \exp(\beta_q \times (u_f - p_{f,r'} - c_{f,i,j'}))} \tag{4}$$

where β_q is inverse standard deviation of the Gumbel distribution of $w_{q,f,r}$. If $\beta_q = \beta$ for all consumers and asking price of product type f is equal for any retailers in location j, eq. 4 is simplified into:

$$P_{i,j} = \frac{\exp(-\beta \times (p_j + c_{i,j}))}{\sum_{j' \in J} \exp(-\beta \times (p_{j'} - c_{i,j'}))} \tag{5}$$

In this simplified case, probability $P_{i,j}$ is only a function of locations i and j. The model of eq. 5 looks similar to a multinomial logit model (McFadden 1980, 2001); however, the source of stochasticity is not modelling uncertainty but taste variation. As such probability $P_{f,q,i,r,j}$ does not reflect estimated likelihood of $[q, i]$ choosing option $[r, j]$, but rather share of $[q, i]$ decisions directed towards (matching) [r,j].

The utility function of eq. 3 implies that each retailer r offers one, and only one, variation of product type f. Therefore, each consumer q associates only one random utility component $w_{q,f,r}$ per retailer, reflecting the fit between this sole product variation and the preferences of the consumer. This assumption is quite unrealistic; typically, a retailer will stock a variety of products of the same type. It is reasonable to assume that larger shops will stock larger varieties. Therefore, variety can be expressed as a function of floorspace (eq. 6):

$$u_{f,q,i,r,j} = \max_{f_v}[(u_{f_v} + w_{q,f_v,r}) - p_{f_v,r} - c_{f_v,i,j}] \tag{6}$$

which means that the utility of consumer q shopping from retailer r is equal to the the utility of buying the specific product variety f_v offered r that maximises the consumers utility. Following Daly and Zachary (1976) and Ben-Akiva and Lerman (1985), and assuming $u_{f_v} = u_f$, $p_{f_v,r} = p_{f,r}$, $c_{f_v,i,j} = c_{f,i,j}$ and $p_r = p$ the probability of consumer q shopping from retailer r is equal to:

$$P_{i,r,j} = \frac{s_r^\alpha \exp(-\beta \times c_{i,j})}{\sum_{[r',j'] \in [R,J]} s_{r'}^\alpha \exp(-\beta \times c_{i,j'})} \tag{7}$$

where s_r is floorspace of r and α is the level of correlation between the stochastic components of the product variations f_v. Macroscopically, eqs. 6 and 7 suggest that the utility that consumer q gets from buying from retailer r is either sub-linear ($\alpha < 1$), linear ($\alpha = 1$), or super-linear ($\alpha > 1$) functions of the floorspace of retail unit r. The respective consumer utilities are translated into either internal dis-economies of scale (sub-linear utility function) or internal economies of scale (super-linear utility function). Indeed, in the former case if a shop doubles its size it will attract less than double its original consumption, in the latter case it will attract more than double its original consumption (eq. 6).

These internal (dis)economies of scale emerge from the application of a transparent utility-based approach and reflect perceived opportunities at the level of individual retailer. The value of α suggests intensity of product variation as a function of floorspace: if $\alpha < 1$ is smaller than 1, the variation of stocked products (as perceived by the consumer) builds up slower than floorspace, if $\alpha > 1$ perceived variation builds up faster than floorspace (e.g. when a shop doubles in size it stocks more than double the number of the original varieties).

Eq. 7 captures the potential impact of size at the individual shop level (internal economies of scale). As such it is sufficient in describing flow patterns and activity distribution associated with shop-size variations. It fails, however, to account for the concentration of retail activity in clusters (markets/shopping centres). In other words, eq. 7 cannot explain agglomeration effects in the spatial distribution of retail activity. In order to introduce the appropriate mechanisms, we extend our behavioural approach accordingly.

Let us assume consumer q evaluating the prospect of shopping a variation of product type f from retailer r. We assume that the consumer knows r's floorspace s_r in advance. As such q can estimate the expected utility of r and decide the probability of visiting r using eq. 7. However we assume that q is only able to determine the stochastic part of utility $w_{q,f_v,r}$ for each of the product variations offered by r upon arrival. In order to minimise the risk of unfavourable matches between r's product variations and personal preferences, q evaluates favourably retailers that are close to other retailers. Therefore, when considering the expected utility of buying product f from retailer r, the consumer also takes into account the expected utility of all other retailers in the vicinity of r. This means that the amount of opportunity associated with r is equal to:

$$S_r = s_r^\alpha + \sum_{r',d_{rr'}<d} s_{r'}^\alpha \exp(-\gamma \times d_{r,r'}) \tag{8}$$

where, $d_{r,r'}$ is some form of generalised distance between r and some other retail unit r' within distance d from r; $s_{r'}$ is the floorspace of r' and γ is a calibration parameter associated with the type of generalised distance used. Since $\exp(-\gamma \times d_{r,r}) = 1$, eq. 8 can be simplified by taking s_r^α inside the sum and widening the summation condition, so r' can be equal to r. Eq. 8 states that the composite perceived utility S_r that a consumer attaches to a particular retailer r is equal to its individual utility s_r^α plus the utility of reaching, from r, the shops in its vicinity. By replacing eq. 8 into eq. 6 we get:

$$P_{i,r,j} = \frac{\exp(-\beta \times c_{i,j}) \sum_{r',d_{rr'}<d} s_{r'}^\alpha \exp(-\gamma \times d_{r,r'})}{\sum_{[r'',j''] \in [R,J]} \exp(-\beta \times c_{i,j''}) \sum_{r',d_{r''r'}<d} s_{r'}^\alpha \exp(-\gamma \times d_{r'',r'})} \tag{9}$$

In the resulting location choice model, α controls the extent of the internal economies and γ of the external economies of scale. If $\alpha > 1$, the relationship between consumer perceived utility of a shop and its size is super-linear and the economies of scale are positive. Similarly, if $\gamma \to +\infty$, the utility of visiting retailer r is only a function of its own utility s_r^α, while if $\gamma \to +0$, the utility of visiting r is shaped equally by all shops r' in the vicinity of r. Note that when $\gamma \to +\infty$ and $\alpha = 1$, we get the simple multinomial logit model. It is easy to show that the model of eq. 9 is a particular case of the Cross-Nested Logit model (CNL) (Wen and Koppelman, 2001; Bierlaire, 2006), in which each retailer r represents a nest, the extend of overlap between nests is controlled by $\exp(-\gamma \times d_{r,r'})$, and there is no correlation between the random elements of utility of individual retailers. At one hand, this observation confirms the proposed behavioural explanation of the agglomeration effect (consumers hedging probability of finding a variety that meets their profile). At the other hand, it offers an alternative explanation; one where consumers prefer destinations with high spatial concentration of retailers with uncorrelated product varieties, in order to bundle together heterogeneous shopping activities (e.g. Oppewal and Holyoake, 2004). While this is also an attractive interpretation of observed agglomeration effects, it fails to account for concentration of retailers offering similar product varieties - a frequently observed phenomenon. In any case, identification of proper interpretation of the underlying behavioural processes is outside the scope of this paper and subject to further work following the development of the model and access to appropriate data sources.

3.2 Implementation Strategy

The aim of this modelling process is to explore how the sizes of competing shopping destinations affect the locations consumers decide to shop from. In section 3.1 we formulated a location choice model for consumers using random utility theory. The proposed cross-nested logit model designed has been designed to capture internal and external economies of scale. In this particular context, economies of scale are defined as the consumers super-linear utility returns with retail size; i.e. the consumers preference to shop at larger shops (internal economies of scale) and at locations with higher concentration of retail activity (external economies of scale).

Using the model in its general form (eq. 9) we can estimate the turnover of a particular retailer r as the sum of sales to all consumers:

$$Y_r = \sum_n \sum_i (X_{n,i} \times P_{n,i,r,j}) \tag{10}$$

where $X_{n,i}$ is the disposable retail budget for consumer of type n based in location i.

Based on the assumption of monopolistic competition and low entry/exit costs to the market, we expect the ratio of turnover to floorspace rent $(fracY_rR_r)$ to be constant for all retail units r of type s. This means that we consider long-term selling prices and running/labour costs per square metre of floorspace to be equal across space (a reasonable simplification for an intra-urban retail market).

To simplify we have assumed that all retailers are of a sole type s. Moreover, we consider only two consumer types (consumers shopping from (i) home and (ii) work) and set a fixed disposable retail budget $X_{n,i} = X$ for all consumer types n and origins i.

Following these simplifications, it is possible to evaluate the level of correlation between estimated turnover Y_r of retailer r (from equations 9 and 10) and some sort of proxy rent R_r for different values of α and γ; and to evaluate the respective internal and external economies of scale.

4 Results

As we mentioned in section 3.2, calibration of the model is a two-step process. In the first instance we use the London Travel demand survey to generate distance profiles for shopping trips. Once these are established we generate location choice patterns for a spectrum of internal and external economies of scale and for each combination we correlate total number of trips to each shop with its rateable value.

4.1 Distance profiles

For this paper we focus on two types of trips: shopping trips from home and work. The LTDS database contains 5004 trips between home and shopping and 2242 trips between work and shopping. These numbers are not sufficient to calibrate the location choice model directly. In fact the LTDS supplementary report (TfL 2011) suggests that the sample size is sufficient only for a Inner/Outer London spatial classification. To address this the survey is only used to calibrate distance profiles; i.e. the probability distribution of distance travelled for shopping from home and work.

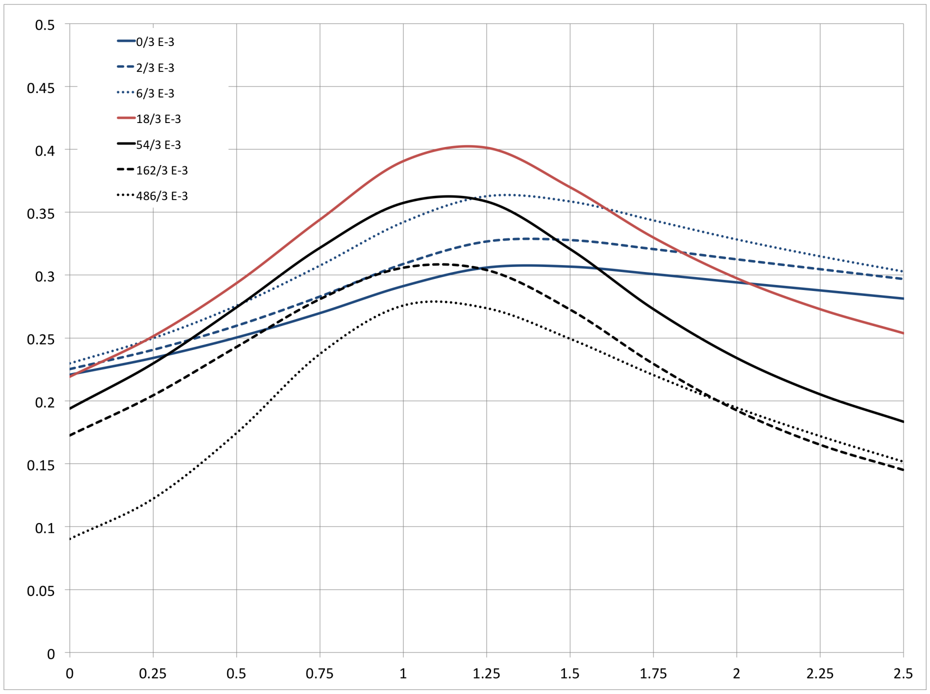

Fig. 1. Correlation between modelled turnover/sq.m. and observed (VOA Rateable Value) floorspace Rent/sq.m.

The calibration of the distance profiles involves three steps: (i) for each $[\alpha, \gamma]$ combination, calculate the attractiveness of every shop; (ii) for each trip origin (home/work locations) calculate distances to every shop (iii) for each $[\beta, \lambda]$ combination generate the cumulative trip distance profile for home and work based trips using eq.11.

$$P_{[type]}(x) = \frac{1}{\sum_i X_i} \sum_{i \in [type]} X_i \sum_{r, d_{ir} < x} \frac{s_r \times exp(-\beta \times f(i,r))}{\sum_r s_r \times exp(-\beta_t \times f(i,r))} \qquad (11)$$

where s_r is the attractiveness of shop r, X_i is the demand for shopping in location i of type $[type]$ (from home or from work) and $f(i,r) = \frac{x^{\lambda}-1}{\lambda}$ is the cost function for origin-destination pair $[i,r]$. Eq. 11 states that the probability of shopping within distance x from origin i is equal to the weighted sum of probabilities of shopping to any shop that is within distance x from i. For each $[\alpha, \gamma]$ combination of the cumulative distribution of eq. 11 we calculate the $[\beta, \lambda]]$ values that maximise the likelihood of the observed LTDS trips. The output of this process generates the following (β, λ) values for home-based and work-based trips respectively: (1.20, 0.42) and (1.34, 0.20).

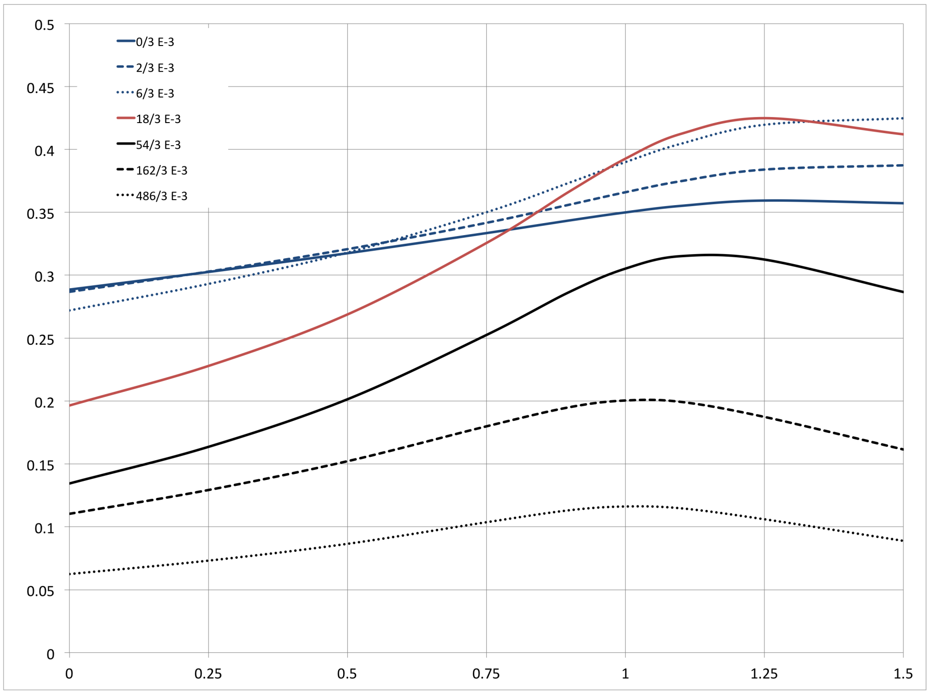

Fig. 2. Correlation between modelled turnover and approximated number of visits (Foursquare check-ins in immediate vicinity).

4.2 Generation of economies of scale

Having completed the calibration of the distance profiles we calculate modelled turnover estimates for each retailer r for a set of (α, γ) parameters. Following this, for each (α, γ) combination, we calculate the correlation level between the modelled turnovers and the observed floorspace rents. For each retailer r, we use the VOA rateable value as an indicator for willingness to pay for floorspace (R_r).

As we mention earlier, rateable value represents the Valuation Office Agencys 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; and as such, it is considered a very good indicator of the property value of the respective hereditament. Figure 1 shows correlation between modelled Y_r and observed R_r for different pairs of values α and γ.

Pair ($\alpha = 1.00, \gamma = 486/3 \times 10^{-3}$) is the combination closer to the simple logit model ($\alpha = 1$, $\gamma = +\infty$); in this case weighted correlation between floorspace rents and revenues is 0.275. On the other hand, correlation is maximum for the pair ($\alpha = 1.25$, $\gamma = 18/3 \times 10^{-3}$). In this case both internal and external economies of scale are manifested. A value of α over 1.00, means that there is a super-linear relationship between the floorspace area of a retail unit and its perceived utility. Similarly, a value of $\gamma << \infty$ means that consumers associate the composite utility of shopping in a shop as a combination of its individual utility and the utility of shopping in shops in its vicinity. As expected this generates strong agglomeration effects that are translated into retail activity clustering and reflected in higher floorspace rents in locations of higher concentration of shops.

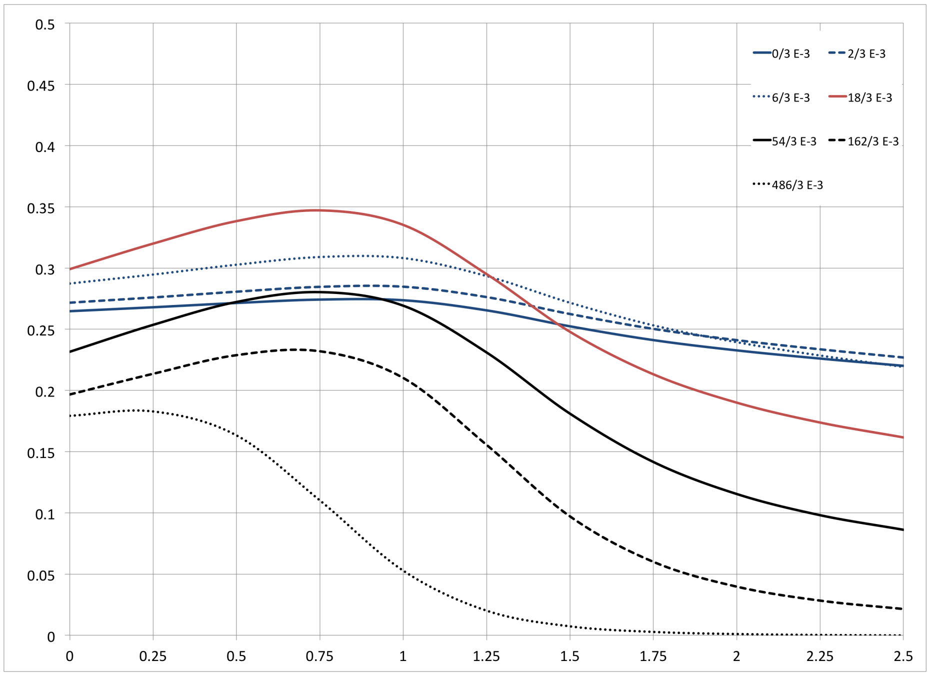

Fig. 3. Correlation between estimated local clustering level and observed (VOA Rateable Value) floorspace Rent/sq.m.

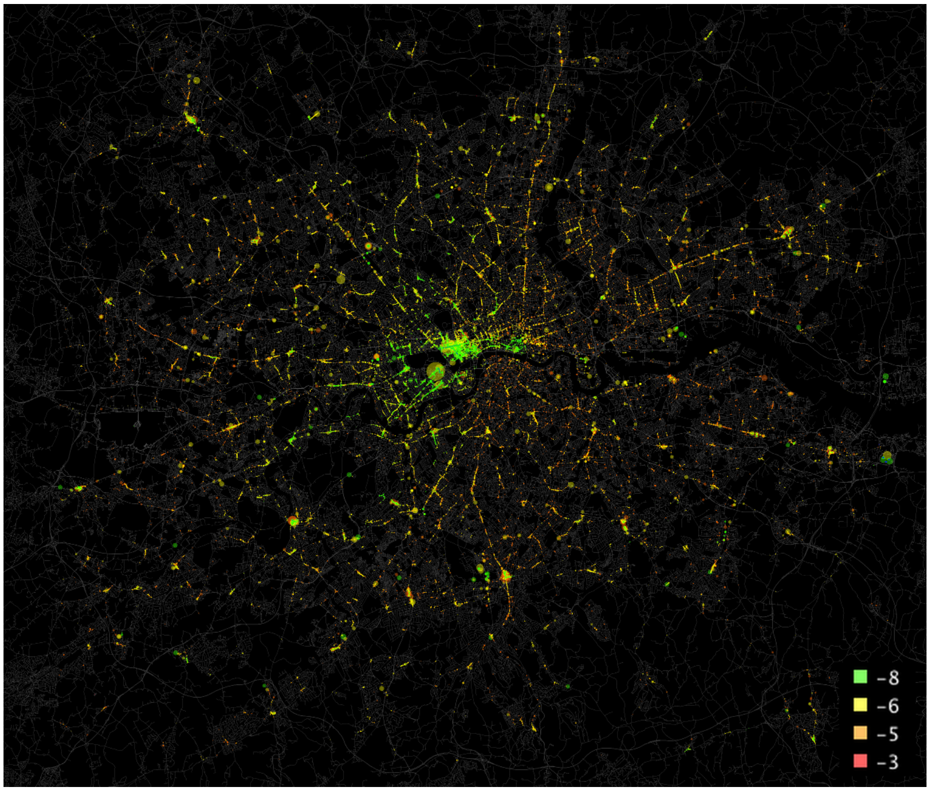

Fig. 4. Log-ratio of modelled turnover to observed floorspace rent for values pair ($\alpha = 1.00$, $\gamma = 486/3 \times 10^{-3}$).

In order to validate the findings of the fitting process presented in figure 1 we use the Foursquare check-ins dataset described in the datasets section as a benchmark. For each retailer r, we sum the number of check-ins in its immediate vicinity (50-100 metres) and correlate modelled turnover (calculated using the process we describe in the methodology section) against the number of check-ins. The results are presented in figure 2. The correlation level of the basic logit model (with no internal or external economies of scale) is quite low at 0.12. On the other hand, maximum correlation of 0.43 is obtained for the exact same pair of economies-of-scale parameters (α, γ) that maximised the correlations in figure 1. This means that the maximum correlation between modelled turnover/sq.m. and VOA floorspace rent corresponds to the same (α, γ) parameters that maximise correlation between modelled turnover and foursquare check-ins. This is quite reassuring, especially if we consider that modelled turnover represents number of consumer visits and forsquare check-ins represent human presence passively collected via social media activity. Therefore, we end up suggesting that the modelled levels of internal and external economies of scale, as perceived by the consumer, that offer the best fit to observed floorspace rents (VOA) and estimated human presence (Foursquare) are, as should be the case, identical and equal to ($\alpha = 1.25$, $\gamma = 18/3 \times 10^{-3}$). This means that internal economies of scale are existent and strongly super-linear (preference towards larger shops) and external economies of scale are also strong (γ is small and therefore the perceived composite utility of a shop is largely determined by other shops in its vicinity).

To establish the impact of modelled consumption demand on observed floorspace rents, we conclude this piece of analysis by calculating the correlation levels between floorspace rents (VOA) and local clustering levels. In this context, local clustering is identified as the accessibility of each shop r to other shops r' as expressed by eq. 8. Essentially, this is equal to considering th case where consumers are uniformly distributed in space, and as such turnover is determined only by the respective perceived utility of each shop. In this case, correlations between clustering levels and floorspace rents are given by figure 3. As expected, correlation levels are lower (since consumer s' spatial distribution is assumed uniform): maximum value is 0.35, and corresponds to the ($\alpha = 0.75$, $\gamma = 18/3 \times 10^{-3}$) parameter pair. Therefore, in this case internal economies of scale are negative (preference for smaller shops), but external economies of scale remain positive and strong at $18/3 \times 10^{-3}$.The fact that the oversimplification of consumer demand has a direct impact on the estimation of the type of internal economies of scale, highlights the importance of developing models that are well-rooted to behavioural attributes of the respective actors. Ideally, all sectoral models should be designed as integrable components of comprehensive spatial equilibrium/dynamics model (e.g. Echenique et al, 2013).

To conclude our discusion, figures 4 and 5 illustrate the log-ratio between modelled revenue and observed floorspace rents ($\log(\frac{Y_r}{R_r})$) for shops. The size of the each circle refers to the floorspace area of the respective shop.

The patterns in figures 4 and 5 highlight the limitations of two assumptions: (i) homogeneity of population in respect to disposable retail budget, and (ii) network-based metric distance as the determinant of proximity between retailers r and r'. In the case of the former, it is clear, from the map in figure 5, that, despite addressing some of the distortions seen in the map of figure 4, there is still systematic overestimation of turnover in the south and east sides of London (where household incomes are relatively low) and systematic underestimation of turnover in the west and south west sides of London (where household incomes are higher than the London average). The introduction of disposable budgets in line with household incomes would, to an extend, address this issue. In the case of the latter, looking at the distribution of log-ratios in figure 5 (particularly in Central London), it becomes clear that there is systematic overestimation of turnover of shops r that are in proximity to other shops r' in terms of metric distance, but not in terms of topological distance; e.g. routes from r to r' are complicated, involving several turns. The introduction of composite metric/topological costs of moving from r to r' should partially address this issue, and the balance between metric and topological cost components is a potential area for further research (Zachariadis, 2014).

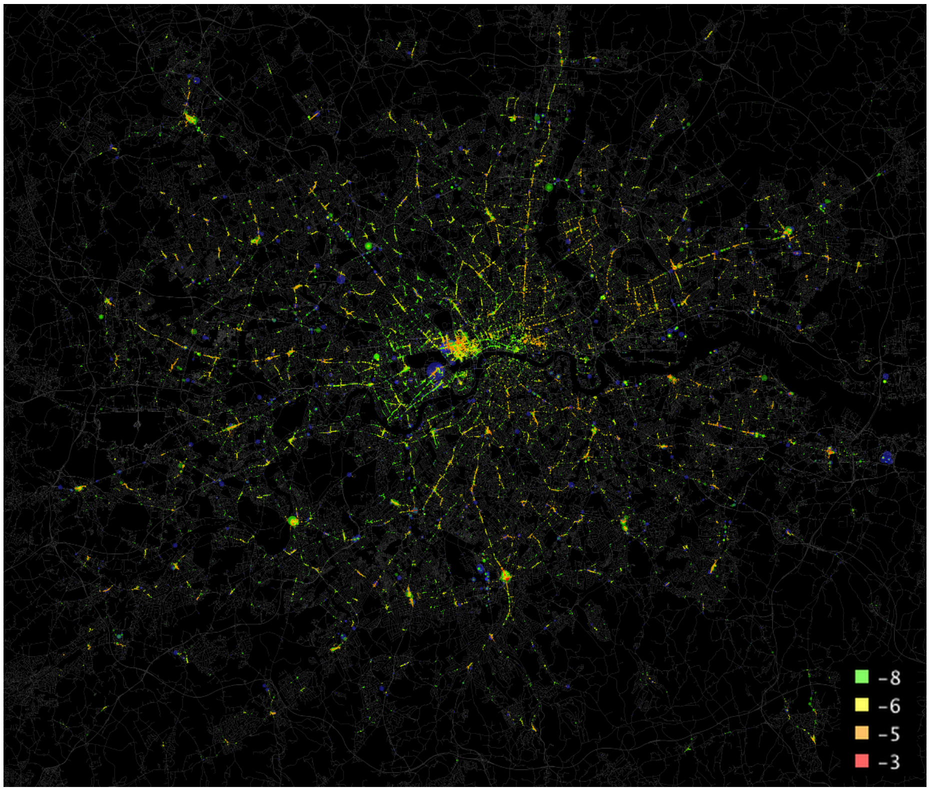

Fig. 5. Log-ratio of modelled turnover to observed floorspace rent for values pair ($\alpha = 1.25$, $\gamma = 18/3 \times 10^{-3}$).

5 Conclusion and Next steps

In this paper we present a location choice model, based on random utility and following the, growing in popularity, cross-nested choice structure (Wen and Koppelman, 2001). The novelty of the proposed model is that it models retailers at the individual level. This opens up exciting opportunities towards integrating the consumer location choice component with explicit retail location microsimulation models able to get full advantage of emerging availability of detailed data sources and incorporate complex behaviour 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 floorspace 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. Business Structure Dataset from ONS) offer sufficient detail to capture all the main dimensions of behavioural variation.

Having said that, 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 translates a substantial amount of the discrepancy between the modelled flows of the unconstrained location choice model and the observed rents into estimates of internal and external economies of scale.

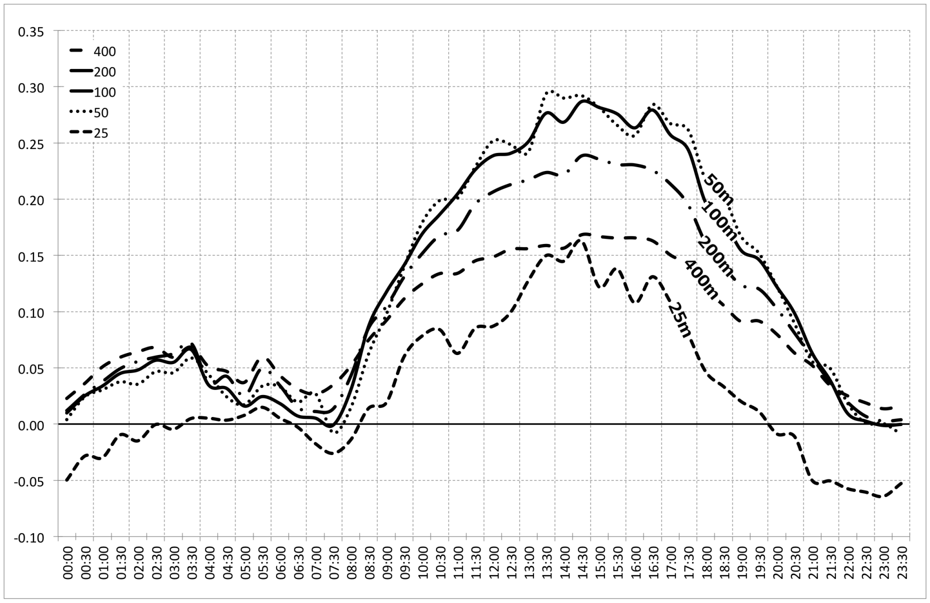

Fig. 6. Correlation between number of Twitter Users and Rent-rate for Retail premises (x-axis: time of day, y-axis: correlation). Each line represents distance from retail premise (e.g. 100m represents correlation between rent rate of shop and number of Twitter users within 100 metres from the shop).

Looking at the - not too distant - future, passively generated datasets of human presence promise to offer deeper insights on the dynamics of urban activities, including spatio-temporal patterns of shopping behaviour. For example, figure 6 illustrates the correlation between number of tweets and rateable value of retail stores (Manley et al, 2015). Lines represent distance from stores; e.g. the 100m line represents the correlation level (y-axis) between rateable value of each store and number of tweets within 100 metres from it for different times of the day (x-axis). Figure 6 shows that the highest correlations between rateable values and number of tweets are found for the 50m and 100m distance bands (there is little difference between the two) and between 2pm and 5pm. Both spatial and temporal dimensions seem to confirm expected values (afternoon shopping and distances that cover in-store locations plus the immediate vicinity of retail environments).

Exercises like this one, are particularly useful for exploring the extent in which biases associated with the temporal variation of social-media usage (i.e. preference of users to tweet at specific times) is reflected in the temporal distribution of generation of digital output associated with particular activities and thus the respective impact on the efficacy of passively generated datasets in generating valid representations of travel demand and activity dynamics.

Having said that, the abundance of existing social media datasources and the relentless pace in which new data are introduced, sustain the promise of accessible and highly disaggregated spatiotemporal information for anyone who manages to overcome lack of specification, representational biases and possibly absence of context.

6 Bibliography

Alonso, W., 1960. A theory of the urban land market. Papers in Regional Science, 6(1), 149-157.

Anderson, S. P., De Palma, A., and Thisse, J. F. (1992). Discrete choice theory of product differentiation. MIT press.

Ben-Akiva, M.E. and Lerman S., 1985. Discrete-Choice Analysis: Theory and Applications to Travel Demand, MIT Press (1985)

Bierlaire, M., 2006. A theoretical analysis of the cross-nested logit model. Annals of operations research, 144(1), 287-300.

Christaller, W., 1966. Central places in southern Germany. Prentice-Hall.

Daly, A.J., Zachary, S., 1976. Improved multiple choice models. In: Proceedings of the Fourth PTRC Summer Annual Meeting. University of Warwick, England, 12-16 July 1976.

Dennis, C., Marsland, D., and Cockett, T., 2002. Central place practice: shopping centre attractiveness measures, hinterland boundaries and the UK retail hierarchy. Journal of Retailing and Consumer Services, 9(4), 185-199.

Dixit, A. and Stiglitz, J., 1977. Monopolistic Competition and Optimum Product Diversity. American Economic Review 67 (3): 297308

Echenique, M.H., Grinevich, V., Hargreaves, A.J. and Zachariadis, V., 2013. LUISA: a land-use interaction with social accounting model; presentation and enhanced calibration method. Environment and Planning B: Planning and Design, 40(6), 1003-1026.

Fujita, M., Krugman, P.R., and Venables, A.J., 2001. The spatial economy: Cities, regions, and international trade. MIT press.

Fujita, M., and Thisse, J.F., 2013. Economics of agglomeration: Cities, industrial location, and globalization. Cambridge university press.

Greenhut, M.L., Norman, G., and Hung, C.S. 1987. The economics of imperfect competition: a spatial approach. Cambridge University Press.

Hotelling, H., 1931. The economics of exhaustible resources. The journal of political economy, 137-175.

Huff, D.L., 1966. A programmed solution for approximating an optimum retail location. Land Economics, 293-303.

Krugman, P., 1998. What's new about the new economic geography?. Oxford review of economic policy, 14(2), 7-17.

Krugman, P., 1990. Increasing returns and economic geography (No. w3275). National Bureau of economic research.

Leszczyc, P.T.P., Sinha, A. and Sahgal, A., 2004. The effect of multi-purpose shopping on pricing and location strategy for grocery stores. Journal of Retailing, 80(2), 85-99.

Manley E., Dennett A., Serras J., Zachariadis V. and Batty M., 2015. Visualising Londons Traffic: Flow and Activity in a 21st Century City. in Traffic in Towns [ed. Jin Y.]. [Forthcoming]

McFadden, D., 1980. Econometric models for probabilistic choice among products. Journal of Business, S13-S29.

McFadden, D. (2001). Economic choices. American Economic Review, 351-378.

Oppewal, H. and Holyoake, B., 2004. Bundling and retail agglomeration effects on shopping behavior. Journal of Retailing and Consumer Services, 11(2), 61-74.

Reimers, V. and Clulow, V., 2004. Retail concentration: a comparison of spatial convenience in shopping strips and shopping centres. Journal of Retailing and Consumer Services, 11(4), 207-221.

Suarez, A., del Bosque, I.R., Rodrguez-Poo, J.M., and Moral, I., 2004. Accounting for heterogeneity in shopping centre choice models. Journal of Retailing and Consumer Services, 11(2), 119-129.

Teller, C. and Reutterer, T., 2008. The evolving concept of retail attractiveness: What makes retail agglomerations attractive when customers shop at them?. Journal of Retailing and Consumer Services, 15(3), 127-143.

von Thnen, J.H., 1966. Isolated state: an English edition of Der isolierte Staat. Pergamon Press.

Train, K.E., 2009. Discrete choice methods with simulation. Cambridge university press.

Wen, C.H., and Koppelman, F.S., 2001. The generalized nested logit model. Transportation Research Part B: Methodological, 35(7), 627-641.

Williams, H.C., 1977. On the formation of travel demand models and economic evaluation measures of user benefit. Environment and planning A, 9(3), 285-344.

Zachariadis, V., 2014. Modelling pedestrian systems (Doctoral dissertation, University College London (University of London)).