A Huff model with firm heterogeneity and selection. Application to the Italian retail sector.*

Massimo Del Gatto[1]
'G.d'Annunzio' University and CRENoS

Carlo S. Mastinu
University of Cagliari

A Huff model with firm heterogeneity and selection. Application to the Italian retail sector.*

Massimo Del Gatto[1]
'G.d'Annunzio' University and CRENoS

Carlo S. Mastinu
University of Cagliari

Abstract

We incorporate firm heterogeneity (in terms of productivity - i.e., marginal costs) in a Huff model of competition in the retail sector. A higher market potential in the trade area is associated to higher average productivity and lower productivity dispersion, through selection of the best stores. The analysis, based on a unique dataset encompassing 14,212 Italian retailers, finds support to this relationship in Southern Italy but not in Northern and Central Italy (where opposite results are obtained in some cases), suggesting the selection dynamics to be affected by context factors (other than provincial/regional accessibility) related to an upper geographical scale. Results are robust to controlling for local context factors such as financial risk and floor size restrictions. Floor size restrictions are found to enhance selection.

Keywords: Huff model, firm selection, accessibility, trade areas, retail location

JEL Classification: R12, F12, R3, L81.

*The authors gratefully acknowledge financial support from the 'Analisi dei costi economici addizionali attribuibili allo stato di insularità, con particolare riferimento alla di differenza rispetto a casi di "geographic remoteness" riconosciuti nellambito della politica regionale europea', a project of European interest funded by the local government of the Sardinia region of Italy [grant number CRP-27162].

[1]Corresponding author. Contact details: Massimo Del Gatto; Università 'G.d'Annunzio', Department of Economics (DEc); Viale Pindaro, 42; 65127 - PESCARA (ITALY); massimo.delgatto@unich.it

1 Introduction

The study of retail trade areas has a long tradition. According to Reilly's (1929, 1931) "law of retail gravitation", the market potential of two competing retail stores depends on their relative size, on the one hand, and their relative distance from potential customers, on the other hand.

Huff (1962, 1963) contributed to the quantification of retail trade areas by modeling the probability that consumers patronize different competing stores within the same area. According to the Huff model, this probability is a function of the store accessibility, relative to its competitors, and can be estimated through gravity regressions. The popularity and longevity of this approach can be attributed to its comprehensibility, relative ease of use, and its applicability to a wide range of problems. Several generalizations have been developed (e.g., Nakanishi and Cooper, 1974 and 1982), but the logic of the model remained basically unchanged. In particular, with the development and diffusion of new methodologies, based on the use of Geographic Information Systems (GIS), the model has been greatly extended and enriched (Birkin, 1995; Satani et al., 1998; Huff, 2003; Suárez-Vega et al., 2011; Li and Liu, 2012) and is now the main tool used by retailers in choosing the location of their stores. However, while the Huff model is regarded as a cornerstone approach in both geographical (e.g., Kwan, 1998) and marketing (e.g., Bell and Tang, 1998; Grewal et al., 2009) literature, it received relatively little attention from economic literature both at the theoretical and empirical level, with only a few notable exceptions such as Davis (2006).

From this point of view, a dimension which is completely neglected is the potential process of firm selection, associated with competition and market size, stemming from differences in productivity across firms/stores. In fact, as advocated by the "New New Trade Theory" (hereafter, NNTT)[2] models, the size of the market can be associated to selection effects stemming from higher

[2] The expression "New Trade Theory" was coined to refer to a strand of international trade literature, pioneered

factor market (e.g., Melitz, 2003) and/or product market (e.g., Melitz and Ottaviano, 2008; Corcos et al., 2012) competition. According to this type of selection, high-productivity firms succeed to charge lower prices for goods of a given quality, or in offering goods of superior quality at given prices. This allows them to gain market share at the expense of the less productive firms. Through this process, the better firms earn handsome profits, the mediocre ones lower profits, and the worst soon disappear, being unable to cover their production costs with revenues.[3] This reasoning finds ample empirical support for the manufacturing sector (see e.g., Pavcnik, 2002; Bustos, 2011), but its relevance for the retail sector has not been documented so far, at least to our knowledge.

To emphasize this aspect, we incorporate heterogeneous (in terms of productivity and, hence, marginal costs) retailers into the Huff (1963) model, to show that when only size and distance, as well as consumer income, matter for demand (demand is inelastic), the process of firm selection is tougher when the market potential in the area is higher (that is, the larger is the total available income that can be reached and the smaller is total size of the competitors). Space is expensive, and less productive firms can only afford a relatively small selling area. The presence of entry costs imposes a threshold, in terms of size. A larger local market, by increasing the profit maximizing size, lowers the level of costs above which firms are not able to serve the market, thereby decreasing both the average and the dispersion of the marginal costs distribution in the munici-

by Krugman (1980) and further developed by Dixit and Norman (1980), Markusen (1981), Helpman (1984) and Helpman and Krugman (1985) among others, focusing on the role of increasing returns to scale and imperfect competition in international trade. While New Trade models successfully explained some key facts in international trade, such as the emergence of intra-industry flows, subsequent literature highlighted additional competition effects: higher competition forces the least productive firms to leave the market (Bernard and Jensen, 1999; Aw, Chung and Roberts, 2000; Clerides et al., 1998) and induces market share reallocations towards the more productive firms (Pavcnik, 2002; Bernard, Jensen and Schott, 2006). Recent theoretical literature accommodated this "selection effect" by enriching the New Trade Theory approach with the assumption that firms are heterogeneous in terms of productivity (i.e., total factor productivity). This generated the class of models [i.e., Bernard et al. (2003), Melitz (2003), Ottaviano and Melitz (2004)], referred to as "New New Trade Theory".

[3] The productivity of firms that generate revenues barely sufficient to cover costs defines the threshold below which it is impossible for a firm to survive in the market. This threshold of survival determines the average productivity of active firms.

pality. Accordingly, trade areas with a higher market potential should be characterized by higher average productivity (i.e., lower costs) and lower productivity dispersion (i.e., lower marginal cost dispersion).

We apply this model by taking advantage of a unique dataset, encompassing information on balance sheet items, size (in square meters) and geographical coordinates for 14212 Italian retailers (hypermarkets, supermarkets, discounters, and small retailers). We compute a theory-based relative measure of market potential (the ratio of distance-weighted consumers' income to distance-weighted store size in the trade area) at the municipality level which, according to the model, should be negatively correlated to the first and second moments of the cost distribution.

Indeed, we find evidence of such relationships in general. Moreover, when we take the geographical articulation (North, Central, South) into account, in order to control a number of factors (e.g., infrastructure, institutional quality and regional autonomy, local regulation, financial institutions, labor market thickness, and human capital, among others) that are likely to affect the effectiveness of the selection process at the local level, we show this evidence to be pervasive in Southern Italy. By contrast, we report mixed evidence for Central and Northern Italy. This evidence is robust with respect to including local factors such as financial risk and entry barriers in the retail sector. Notably, the latter is found to foster the process of store selection.

Our contribution fits into an expanding vein of literature dealing with the competition effects associated with accessibility at the local level. For example, Öner (2017) uses accessibility to study the relationship between place attractiveness and consumption possibilities in rural and city municipalities. Guy (2013) develops a model of competition between walkable shops and shops whose customers drive. De Mello-Sampayo (2016) uses a competing-destinations framework to explain the flows of patients from their residential areas to health supplier regions.

While documenting that the selection effect also takes place at the local level (for non-tradables), our results point out that the effectiveness of the selection process is significantly affected by context factors related to an upper geographical scale. As this aspect has not been highlighted earlier for the retail sector, it might deserve further attention, as the identification of the factors affecting the pervasiveness of the selection process at the local level might provide the public authority with important policy implications. While we leave this issue for further research, we show that the extent of provincial or regional accessibility is not one of those context factors.

The paper is organized as follows. In Sections **??** and **??**, we present the model and the empirical strategy. In Section **??** we describe the data. Benchmark results are reported in Section **??**. A number of robustness checks is reported, together a brief description of the dataset, in the web appendix. Section **??** concludes.

2 Model

In the original Huff model, the probability P_{if} that a consumer i, located in trade area A, will select store f located in site f (letter f is used to refer to both the store and its location), among all possible alternatives in A, is assumed to be a positive function of the store f's sales area (i.e., interchangeably referred to as size hereinafter) s_f and a negative function of its distance τ_{if} from the consumer (indexed i). Size and distance are evaluated in relative terms with respect to all possible alternative stores in A. To represent an alternative, a store has to fall within a given traveling time to the consumer. Hence, the probability can be written as

$$P_{if} = \frac{s_f^{\alpha} \tau_{if}^{\beta}}{\sum_{r=1}^{R(A)} s_r^{\alpha} \tau_{ir}^{\beta}}, \tag{1}$$

where index r identifies a general retailer (i.e., store), and $R(A)$ denotes the number of retailers in A.[4] Since demand is inelastic, the total demand available to store f can be expressed as

$$D_f = \sum_{i=1}^{I(A)} P_{if} B_i = s_f^{\alpha} \Phi_f \qquad \text{with} \quad \Phi_f = \frac{\Psi_f}{\Theta_A}, \quad \Psi_f = \sum_{i=1}^{I(A)} B_i \tau_{if}^{\beta} \quad \text{and} \quad \Theta_A = \sum_{i=1}^{I(A)} \sum_{r=1}^{R(A)} s_r^{\alpha} \tau_{ir}^{\beta} \tag{2}$$

where $I(A)$ is the number of consumers located in A and B_i is their income. As well as on its size, a store f's total demand depends on the two terms Ψ_f and Θ_A. The former is a distance-weighted measure of the total consumer income accessible from location f, which varies across locations within the trade area. The latter is a distance-weighted measure of the total store size in the area and is therefore specific to the area. The ratio of these two terms (i.e., Φ_f) can be referred to as the *Available Market per Unit of Sales Area* (hereafter, AMUSA). This is a relative measure of (within-area) market potential which is comparable across trade areas.

This simple demand structure can be used to contextualize the original Huff model into a framework in which heterogeneous retailers choose their profit maximizing (floor) size. Heterogeneity is expressed in terms of inverse total factor productivity – i.e., Unit Input Requirement (UIR) – hereinafter referred to as c_f.

Retailers' activity only requires land (i.e., the surface area), as a production factor. The production function is modeled as $Y_f = s_f^{\gamma}/c_f$. Firms use the same technology, but differ in the UIR term. Markets (areas) are segmented, entailing that multi-store firms independently maximize profits from different locations, so that the decision concerning store f at location f can be always

[4]The demand function in equation (**??**) is based on the assumption that the probability that consumer i, confronted with a set of alternatives, will select a given store is directly proportional to the perceived utility of each alternative. The choice is probabilistic, and each store is characterized by a positive probability $P_{if} = U_{if}/\sum_r^{R(A)} U_{ir}$ to be chosen, with U_{if} denoting the consumer's utility associated to choice f (and $\sum_r P_{ir} = 1$). Assuming that U_{if} is directly proportional to s_{if} and inversely proportional to τ_{if}, with the degree of proportionality expressed by the two parameters α and β, yields the demand function in equation (**??**). Inelasticity entails that, if only one store existed, the total number of consumers would patronize it regardless of where it is located.

dealt with as the decision of a single-store firm with UIR equal to c_f.[5] Thus, the two terms 'store' and 'firm' can be used interchangeably. The marginal costs faced by firm f is ωc_f, with $\omega = r^{\gamma}$. Here, r denotes the rental price of the surface area, which can be either specific or non-specific to the trade area (in the application we set it as province specific).

Firm-store f sets its size by solving

$$\underset{s_f}{\text{Max}} \quad D_f - \omega c_f s_f \;=\; s_f^{\alpha}\Phi_f - \omega c_f s_f \tag{3}$$

taken as given the total sales area in A. First order conditions yield

$$s_f^* = \left[\frac{\omega c_f}{\alpha \Phi_f}\right]^{\frac{1}{\alpha-1}}. \tag{4}$$

Under the standard assumption that $\alpha < 1$, a negative relationship between optimal store size and the firm's UIR emerges. In fact, the profit maximizing floor size is higher when the UIR is lower (the more productive is the firm), and the AMUSA is higher.

Free entry in the trade area imposes

$$\int_0^{\bar{c}_f} s_f^{\alpha}\Phi_f - \omega s_f c_f \, dG(c) = z_E \omega \tag{5}$$

where $\bar{c}_f$ refers to the cutoff UIR level above which stores are not able to keep serving the market, and $z_E\omega$ is a fixed entry cost that each firm has to bear in order to open a new store. Also, z_E can be thought of as either specific or non-specific to the trade area. In the application, we assume it to be the same for all the Italian municipalities.

[5] Alternatively, one can imagine production to also require labor, with the latter inelastically provided by consumers at unit wage (i.e., $Y_f = c_f^{-1} s_f^{\beta_s} l_f^{\beta_l}$).

In order to solve equation (**??**) explicitly, let us assume, as in Melitz (2003) and Melitz and Ottaviano (2008), that UIR follows a Pareto distribution $G(c_f)$ with shape parameter k and upper bound c_{max} within the support $[0, c_{max}]$: $G(c_f) = (c_f/c_{max})^k$. By determining the probability to observe a store with UIR below a given level, c_{max} can be thought of as a parameter subsuming the exogenous differences in terms of the socio-economic context in which the stores are located. Examples can be the quality of infrastructure and financial institutions, or the local regulation concerning the retail sector. In the application we imagine the support of the UIR distribution to vary across the Italian macro-regions (North, Central, and South-Islands).

Using (**??**) to substitute for s_f, the solution of (**??**) provides us with the following expression for the UIR cutoff above which store f is not able to survive in area A:

$$\bar{c}_f = \Lambda \; z_E^{\frac{\alpha-1}{\upsilon}} \; \omega^{-\frac{1}{\upsilon}} \; (c_{max})^{\frac{k(\alpha-1)}{\upsilon}} \; \Phi_f^{\frac{1}{\upsilon}} \tag{6}$$

where Λ is equal to $\frac{k+1}{k[\alpha^{-\alpha/(\alpha-1)} - \alpha^{1/(\alpha-1)}]}$ and $\upsilon = k(\alpha-1) + \alpha$.[6]

Under the stability condition that $k > \frac{\alpha}{1-\alpha}$ (which implies that $\upsilon < 0$), *the UIR threshold increases with the fixed cost of entry and with the rental price of the surface area, and decreases with the AMUSA.*[7]

The rationale for the AMUSA effect is as follows. From equation (**??**), a higher AMUSA value entails, for all firms, a higher profit maximizing floor size. However, the surface area imposes a rent cost and the profit maximizing size is lower for retailers with a relatively high UIR. By imposing a minimum size, the entry cost also imposes a maximum possible UIR (i.e., UIR cutoff). Retailers with UIR values above this threshold level cannot afford a sufficiently large size (i.e.,

[6]Note that, given the one-to-one relationship between UIR and size, the equilibrium UIR no longer depends on size, once equation (**??**) is used to substitute for s_f into (**??**).

[7]The condition $k > \frac{\alpha}{1-\alpha}$ allows the integral in (**??**) to converge to a meaningful solution.

they have a too low profit maximizing floor size) and are not able to survive.

3 Testable implications and empirical strategy

Equation (**??**) yields testable implications that can be studied through the various moments of the UIR distribution characterizing the trade areas. In fact, with the UIR Pareto distributed, a lower UIR cutoff maps into lower measures of central tendency (e.g., mean and median) and lower measures of dispersion (e.g., standard deviation and interquartile range).[8] In particular, the UIR average and standard deviation in a given trade area are given by

$$\text{AVG}(c_f) = \frac{k}{k+1}\,\bar{c}_f \qquad \text{and} \qquad \text{SD}(c_f) = \left[\frac{k}{(k+1)^2\,(k+2)}\right]^{1/2}\bar{c}_f \tag{7}$$

These two expressions, together with the expression for $\bar{c}_f$ in equation (**??**), reveal the aggregate effect of the selection process featured by the model: a *higher AMUSA* value is associated with a *lower UIR threshold* (through (**??**)) and (through (**??**)) with *lower UIR average and standard deviation* in the area.

Accordingly, the hypothesis that we bring to the data in the following analysis is the existence of a *negative relationships* between AMUSA and both average UIR (i.e., $\partial \text{AVG}(c_f)/\partial \Phi_f < 0$) and UIR standard deviation (i.e., $\partial \text{SD}(c_f)/\partial \Phi_f < 0$).

To this purpose, trade areas are defined at the municipality level, considering all the munici-

[8] The simplest measure of dispersion would be the 'range', defined as the UIR gap between the best- and worst-performing stores. However, given the support $[0, \bar{c}_f]$, the range is simply equal to $\bar{c}_f$, so it increases with the degree of competition in the trade area. While easily understood, being based on the two boundary values only, the range is necessarily very sensitive to extreme observations and should be used together with other measures. The 'standard deviation' is the most widely used measure of dispersion. Although less sensitive, the SD might also be problematic in highly skewed distributions. In the web Appendix we provide robustness checks for both average and dispersion, by relying on the median and the interquartile range.

palities accessible in a given traveling time from the given municipality (which amounts to setting locations f equal to the municipality). We first use information at the single store level to estimate each store's UIR and then aggregate at the municipality level obtaining, for each municipality, the AVG and SD of the estimated UIR distribution (that is, without estimating k and $\bar{c}_f$). Accordingly, we compute the AMUSA for each municipality as the ratio of the distance-weighted total income in the trade area (i.e., the distance-weighted income of all the municipalities accessible from the given municipality) to the distance-weighted total store size in the trade area. For the latter, we first compute the total sales area in each municipality and then sum, for each municipality, over the sales area of all the accessible municipalities, weighting by distance.

Under this strategy, our estimating model is obtained using the expression in (**??**) to substitute for the UIR cutoff $\bar{c}_f$ in (**??**), and taking logs:

$$\begin{aligned} ln(Q) &= \lambda_0 \ + \lambda_1 ln(AMUSA) \ + \lambda_2 ln(\omega) \ + \lambda_3^N North \ + \lambda_3^S South \ + \\ & \quad + \lambda_4^N ln(AMUSA) * North \ + \lambda_4^S ln(AMUSA) * South \ + \epsilon. \end{aligned} \tag{8}$$

Depending on the specification, the dependent variable Q is the UIR average (AVG) or the UIR standard deviation (SD) at the municipality level (in the web appendix we also present results for the median and the interquartile range). The rental price of the surface area ω is assumed (see Section **??**) to be province-specific. AMUSA is, as said, municipality-specific. $North$ and $South$ are dummies for being located in Northern and Southern Italy, respectively (Central Italy being the excluded category). ϵ is an iid error term.

If we imagine each store's UIR as a draw from a common UIR distribution, c_f is a random variable over the support $[0, c_{max}]$, with c_{max} subsuming the ex-ante probability of 'good' UIR draws: the lower c_{max}, the higher is the probability of drawing a low enough UIR (that is, a UIR that will allow store f to survive the selection process). In principle, as the model is conceived at the level of the trade area, the empirical application should allow stores located in different trade areas to draw from different distributions (one for each trade area). However, this is not possible, given the available data. We tackle this issue by estimating different specifications of equation (**??**). In our fist specification, we include only AMUSA and rental price of the surface area as regressors, and estimate λ_0, λ_1 and λ_2 under the hypothesis that all the Italian stores face the same ex-ante UIR distribution. In this case, as well as capturing the effect of z_E and Λ, the estimated λ_0 encompasses the UIR upper bound c_{max} (the same throughout Italy). Under this specification (columns 1 and 4 in Tables **??** and **??**), the parameter of interest, capturing the average AMUSA effect in Italy, is λ_1. The expected sign of λ_1 is negative.

It is well-known that the Italian territory is characterized by huge differences in socio-economic features (e.g. infrastructure, local regulation, financial institutions, labor market thickness and human capital, among others). By interacting with local competition and market size (our variables of interest), these factors are likely to impact the effectiveness of the selection process. The model suggests a convenient way to deal with these factors: we can imagine them to condition the UIR distribution by affecting its upper bound c_{max}. Since most of the heterogeneity in these characteristics is correlated with latitude, their action can be easily taken into account by including a vector of dummies controlling for municipality belonging to one of the three Italian macro-regions: North, Central and South (with the two island regions, Sardinia and Sicily, included in the South). Under this specification, the UIR distribution varies across macro-regions, so that the dummy effects can be traced back to exogenous differences in the UIR upper bound c_{max}, as well as in z_E and Λ. Under this specification (columns 2 and 5 in Tables **??** and **??**), the estimated $\lambda_0, \lambda_3^N, \lambda_3^S$

capture these effects in Central, Northern and Southern Italy respectively (Central Italy is the omitted category), with λ_1 still representing the average AMUSA effect in Italy.

Finally, as well as conditioning the ex-ante UIR distribution, geography might also affect the effectiveness of the selection process. In order to isolate the differential effect of the AMUSA in the different macro-regions, we estimate a further benchmark specification including the interaction terms $AMUSA * North$ and $AMUSA * South$.[9] In this case (columns 3 and 6 in Tables **??** and **??**), λ_1 represents the average AMUSA effect in the Center, while $\lambda_1 + \lambda_4^N$ and $\lambda_1 + \lambda_4^S$ capture the average AMUSA effect in Northern and Southern Italy, respectively. According to Equation (**??**), the expected sign of λ_1, $(\lambda_1 + \lambda_4^N)$ and $(\lambda_1 + \lambda_4^S)$ is negative in both the AVG and SD regressions.

4 Data and variables definition

Our main explanatory variable is the AMUSA. To compute this term, we use GIS software to calculate Ψ_f and Θ_A by aggregating over consumers, for Ψ_f, and stores, for Θ_A, within a traveling time of fifteen minutes. For each municipality, the trade area A consists of all the municipalities located within such traveling time.

Data on retailers are provided by Nielsen, which conducts a regularly updated census on Italian mass retailers.[10] Our data, updated in September 2016, include 27,966 stores divided into four categories: hypermarkets, supermarkets, discounters, and small retailers. For each store, in

[9]In principle, one might want to go deeper in the geographical disaggregation and estimate region-specific AMUSA effects and UIR distributions. Although we explore this dimension in the web appendix, this exercise is beyond the scope of our benchmark analysis. Since we are concerned with providing 'general results' concerning the effectiveness of the selection effect, our analysis requires detailed data, on the one hand, as the reference unit is the retail trade area, and a wide geographical scale, on the other hand, in order to avoid results that are too much specific. Moreover, a higher number of observations would be needed in order to obtain consistent estimates at a regional level. This is particularly true in the dispersion analysis, where (see Tables **??** and **??**) the number of observations shrinks considerably because of the municipalities in which only one store is observed.

[10]See: http://www.nielsen.com.

addition to geographical coordinates, a number of variables are provided, including the size of the sales area, expressed in square meters. A map with the geographic distribution of stores is reported in the web appendix.

Since c_f is defined as inverse tfp, the UIR index of each store is obtained through standard tfp estimation. Assuming a Cobb-Douglas production function for store f (i.e., $Y_f = tfp_f\, K_f^{\alpha} L_f^{\beta}$), tfp_f is recovered by estimating the vector $(\hat{\alpha}, \hat{\beta})$ using the Olley-Pakes procedure (in order to keep simultaneity into account)[11] and deriving $\widehat{tfp}_f$ as the log difference between observed and predicted output (i.e., Solow residual). Capital is measured through the book value of tangible fixed assets[12] and labor is measured through employment. Data are drawn from the AIDA – Bureau van Dijk database.

The estimated coefficients amount to 0.301 (with standard error 0.0104) for capital and 0.699 (with standard error 0.0189) for labor. To this aim, we match the Nielsen data with balance sheet data drawn from the AIDA database. 14,212 (of the 27,965) stores were matched successfully.[13]

For the term τ, an Origin–Destination (OD) matrix among Italian municipalities is needed. This is calculated using the entire network of the Italian extra-urban roads, updated to 2016. The driving times are estimated through a GIS program and by taking into account four key variables: length, direction of travel, hierarchy of the functional road classes, and journey speed. In order to determine the journey speed in each class of road, we referred to the Ministerial Decree of

[11] Simultaneity arises because information on tfp_{ft}, although unknown to the econometrician, is commonly used by the firm in its decision concerning the amount of inputs. This issue makes the error term in the estimation correlated with capital and labor, and the OLS–estimated $(\hat{\alpha}, \hat{\beta})$ biased. The solution suggested by Olley and Pakes (1996) exploits the idea that investment (i.e., the 'proxy-variable') reacts to the changes in tfp observed by the firm and is therefore a function of it. Under reasonable assumptions, this function is invertible and its inverse can be plugged in the estimating equation before proceeding to estimate the production function parameters (see Del Gatto et al., 2011 and Van Beveren, 2012). Although we use ours, the Olley-Pakes routine is implemented in Stata under the command 'opreg' (see Yasar et al., 2008).

[12] The production function $Y_f = s_f^{\gamma}/c_f$ stated in Section **??** is nested in this standard specification as far as stores' sales area is included in the book value of capital.

[13] In the case of multi-store firms, the UIR refers to the main branch.

November 5, 2001 (so-called Decreto Lunardi), which identifies 14 types of roads and assigns to each type a lower and an upper speed limit. We have taken the latter as the reference speed of travel.[14] Our OD matrix includes 8085 Italian municipalities and consists of the driving times that separate each municipality from the municipalities located within a travel time of up to 15 minutes. An alternative travel time area of 20 minutes has been used, finding no notable differences in the econometric results.

With the OD matrix in our hands, we calculate Ψ_f following the potential accessibility formulation proposed by Wegener et al. (2002): $\Psi_f = \sum_j^{F(A)} B_j \, exp(-\rho\tau_{jf})$. Where index f refers to the given municipality, index j refers to the generic municipality located in the area (with F(A) denoting the number of municipalities located within the driving distance of fifteen minutes from f), and τ_{jf} is the vector of journey time between municipality f and municipality j. Furthermore, B_j is the total available income in municipality j, drawn from ISTAT (the Italian National Institute of Statistics). Also, ρ is a decay parameter set at 0.05. The set of the τ_{jf} for all the municipality pairs forms the OD matrix. A map reporting the computed values of Ψ_f is included in the web appendix.

A similar logic is followed to compute Θ_A, with municipalities' income replaced by the total sales area (expressed in square meters) available there, provided by Nielsen. To measure the rental price of the surface area (r_A), we rely on data from the real estate market and consider the average price of sales and rents, downloaded (in September 2016) from a popular Italian real estate website (www.immobiliare.it), at the Nuts 3 level. We also make use of a measure of entry barriers in the retail sector, calculated at the Nuts 3 level (i.e., Italian provinces) by Schivardi and Viviano

[14]Notice that the resulting OD matrix underestimates actual travel times, for different reasons. First, our data only includes the extra-urban roads, so we do not consider the time required to reach the extra-urban road network. Second, the analysis excludes any kind of barriers (such as traffic lights and toll gates). Third, we use the maximum allowed speed as reference speed of travel.

(2011). The index is computed as the ratio of population to admissible floor space (PAFS), as regulated by law.[15] The higher this ratio, the greater the entry restrictions. Finally, we also use an index of financial risk, computed at the Nuts 3 level (i.e., Italian provinces) by ISTAT as the ratio of non-performing to performing loans granted to all types of firms. Descriptive statistics for the main variables used in the analysis are reported in the web appendix.

5 Results

Table **??** reports the results of the benchmark estimation of equation (**??**) for the central tendency (i.e., AVG(c)) effect and the dispersion effect (i.e., SD(c)), in columns 1 to 3 and 4 to 6, respectively.[16]

As an overall effect, the first row suggests a negative relationship between AMUSA and both the UIR average (column 1) and standard deviation (column 4). This is in line with the selection effect predicted by the model (note that both one-sided and two-sided tests are performed). In columns 2 and 5, we control whether the municipality belongs to Northern, Central, or Southern Italy using the *North* and *South* dummies and setting the Central as the benchmark, so that the coefficients of the two dummies (λ_3^N and λ_3^S) represent the differential AVG and SD effects in Northern and Southern Italy. In line with common sense concerning the productivity gap of Southern Italy, the UIR distributions in the South are characterized by higher average and dispersion. In columns 3 and 6, the AMUSA is interacted with the macro-region dummies *North* and *South*. As Central

[15] As noticed by Schivardi and Viviano (2011): the Italian retail sector, which has a prevalence of traditional small stores, underwent a major regulatory change in 1998. A central feature of the new law is that it delegates the regulation of entry of medium-large stores to local authorities. As it turns out, local regulations differ substantially in their approach to competition: in particular, most regions have established stringent ceilings to the floor space that can be authorized for entry of medium-large stores at the local level.

[16] Note how the number of observations differs across the two sets of regressions, due to the presence of municipalities exhibiting zero dispersion (only one store observed), turning into missing in the log-estimation.

Italy is the benchmark case, the coefficient λ_1 represents the average AMUSA effect in the Central municipalities, while the coefficients of $AMUSA * North$ and $AMUSA * South$ (λ_4^N and λ_4^S, respectively) pick the differential AMUSA effect in the Northern and Southern municipalities (with respect to the Central municipalities). Interestingly, we find that while the UIR gap is confirmed, most of the average AMUSA effect estimated throughout Italy (i.e., the AMUSA coefficient in columns 1-2 and 4-5) has to be attributed to the Southern municipalities; in fact, $\lambda_1 + \lambda_4^S$ is negative and significant, while coefficients λ_1 and $\lambda_1 + \lambda_4^N$ are positive, although not significant in both the one-sided and the two-sided tests (bottom of table).

Thus, the evidence in favor of a selection effect fostered by the local market potential is pervasive in the Southern municipalities and actually absent in the rest of Italy. This difference points to the existence of macro-regional characteristics affecting the pervasiveness of the selection effect at the local level. This is something that has not been highlighted before.

The differences in the productivity dynamics (both firm-level and aggregate) across the Italian regions are well known and related to a number of factors. While examining the details of these factors is definitively beyond the scope of the present analysis (see however Calligaris et al., 2016 for a recent analysis), a dimension that is worth considering is analyzing whether the different documented effects are explained by additional competition effects taking place at a 'less local' spatial scale: the province and/or the region. To investigate this, we recompute the AMUSA at the level of the 103 Italian provinces using, for the numerator Ψ_f, the measure of multi-modal accessibility provided by the European Spatial Planning and Observation Network (ESPON) for the Italian provinces. The recomputed AMUSA (Φ_p) is included, together with the AMUSA computed at the municipality level (correlation is 0.0330), in the regressions reported in columns 1 and 4 of

Table ??.[17] Although significant, the provincial AMUSA effect does not absorb the significance of the differential effect in Southern Italy (the interaction term $AMUSA * South$, still considered at the municipality level, remains significant). To check for the same effect at the regional level, we use, in columns 2 and 5, the ESPON measure of multi-modal regional accessibility.[18] Neither the overall selection nor the differential effect characterizing the municipalities located in Southern Italy can be explained by the regional scale of accessibility.

Finally, we ask whether the documented selection process characterizing Southern Italy changes if local context factors are included in the analysis. To this aim, we perform again the regressions in Table ?? including two province-level characteristics. First, the PAFS index of restriction on admissible floor space, as regulated by the Italian law: the higher this ratio, the greater the size restrictions. Second, the degree of financial risk measured through the incidence of non-performing loans (i.e., the ratio of non-performing to performing loans): the higher this index, the more expensive is capital at the level of the local financial system. The output of these regressions is reported in columns 3 and 6 of Table ??. Size restrictions significatively contribute to the selection process by lowering both the first and the second moments of the marginal cost distribution. Financial risk displays a mild positive correlation with productivity dispersion. The evidence on the selection effect associated to AMUSA is unaffected.

It is worth noting how, also in the robustness checks of Table ??, it is in Southern Italy that the selection effect associated with our AMUSA index is pervasive. In fact, the AVG effect tends

[17] The ESPON accessibility measures used in Table ?? are province-level (upper panel) and region-level (bottom panel) variables computed on 2006 data. The accessibility of province/region j is defined as $Acc_j = \sum_r Z_r exp(-\beta \bar{c}_{jr})$, where $\bar{c}_{jr}$ refers to the aggregation, over transport modes (i.e., air, rail, road), of the cost (c_{jrt}) of reaching r from j using transportation mode t - i.e., $\bar{c}_{jr} = -(1/\lambda) ln \sum_t exp(-\lambda c_{jrt})$, where Z_r is GDP-PPS per capita and population in region r, respectively, for the two measures computed at the province and region level, and λ is a parameter indicating the sensitivity to travel cost. The interpretation is that the accessibility of j increases with the number of "accessible" provinces/regions and with their size (either GDP or population).

[18] In this case, we do not divide by Θ_A.

to be positive in Central Italy, although not significant in the first two exercises (columns 1 and 2) and only weakly significant in the third one (column 3). Also in the North, we obtain a weakly positive effect on SD in column 5 (as well as a positive but not significant effect on AVG UIR in column 3).

6 Conclusions

According to NNTT models, a higher market size, through its induced competition effects, allows high-productivity firms to gain substantial profits at the expense of the less productive firms, which are forced to leave the market being unable to cover their production costs with revenues. While this reasoning finds ample empirical support for tradable goods (i.e., the manufacturing sector), its relevance for non-tradables (e.g., the retail sector) has not been documented so far.

To emphasize this aspect, we incorporated heterogeneous retailers into the original Huff (1963) model, the cornerstone model in retail location analysis. Since demand is inelastic, only size and distance, as well as consumer income, matter for demand. On the one hand, a larger floor size imposes higher costs and less productive firms can only afford a relatively small sales area; on the other hand, entry costs impose a threshold, in terms of size. A larger local market, by increasing the profit maximizing floor size, lowers the level of costs above which firms are unable to keep serving the market, thereby decreasing both the average and the dispersion of the marginal costs distribution in the municipality.

We applied this concept to data taking advantage of a unique dataset encompassing 14,212 Italian retailers (hypermarkets, supermarkets, discounters, and small retailers). We computed a theory-based relative measure of market potential (the ratio of distance-weighted consumer income to distance-weighted store size in the trade area) at the municipality level. According to the model, this measure should be negatively correlated to the first and second moments of the cost distribution, through the selection of the best stores.

As well as providing overall evidence of such a relationship in Italy, we took a geographical perspective (i.e., belonging to Northern, Central or Southern Italy) to control for factors (e.g., infrastructure, institutional quality, local regulation, financial institutions, labor market thickness and human capital, among others) that are likely to condition the effectiveness of the selection process at the local level. We found that the evidence concerning the local selection effect is pervasive in Southern Italy and absent in the Northern and Central municipalities (where the estimated relationship is not significant in general and even opposite with respect to the model, in some cases). This evidence is robust with respect to including local factors such as financial risk and size restrictions in the retail sector. Notably, higher size restrictions are associated with tougher selection.

These findings are new in retail literature and suggest that: i) a higher market potential at the retail trade area level can be associated to aggregate productivity advantages, fostered by a process of firm selection; ii) the selection process at the local level is significantly affected by context factors, external to the firms, which are related to an upper geographical scale. While the identification of these factors might provide public authorities with key policy messages, a suggestion we leave for future research, we show that neither regional nor provincial accessibility is among the latter.

References

Aw, B., Chung, S. & Roberts, M. (2000). Productivity and turnover in the export market: micro-level evidence from the republic of Korea and Taiwan (China), *World Bank Economic Review*, 14, 65-90.

Bell, D. R., Ho, T. H., & Tang, C. S. (1998). Determining where to shop: fixed and variable costs of shopping. *Journal of Marketing Research*, 352-369.

Bernard, A. & Jensen, B. (1999). Exceptional exporter performance: cause, effect, or both?, *Journal of International Economics*, 47, 1-25.

Bernard, A., Eaton, J., Jensen, J. & Kortum, S. (2003). Plants and productivity in international trade, *American Economic Review*, 93, 1268-1290.

Bernard, A., Jensen, B. & Schott, P. (2006). Trade costs, firms and productivity, *Journal of Monetary Economics*, 53, 917-37.

Birkin, M. (1995). *Customer targeting, geodemographics and lifestyle approaches*. In P. Longley, & G. Clark (Eds.), GIS for business and service planning. Cambridge, U.K: Geoinformation International.

Bustos, P. (2011). Trade liberalization, exports, and technology upgrading: evidence on the impact of MERCOSUR on Argentinian firms. *American Economic Review*, 101, 304-340.

Calligaris S., M. Del Gatto, F. Hassan, G. Ottaviano and F. Schivardi (2016). Italy's Productivity Conundrum. A Study on Resource Misallocation in Italy. European Commission. *European Economy Discussion Paper*, 030.

Clerides, S., Lach, S. & Tybout, J. (1998). Is learning by exporting important? Micro-dynamic evidence from Colombia, Mexico, and Morocco, *Quarterly Journal of Economics*, 113, 903-947.

Corcos, G., Del Gatto, M., Mion, G. & Ottaviano, G.I.P. (2012). Productivity and firm-selection: quantifying the new gains from trade, *The Economic Journal*, 122(561), 754-798.

Davis, P. (2006). Spatial competition in retail markets: movie theaters. *The RAND Journal of Economics*, 37(4), 964-982.

De Mello-Sampayo, F. (2016). A spatial analysis of mental healthcare in Texas. *Spatial Economic Analysis*, 11(2), 152-175.

Del Gatto, M., Di Liberto, A., & Petraglia, C. (2011). Measuring productivity. *Journal of Economic Surveys*, 25(5), 952-1008.

Dixit, A., & Norman, V. (1980). *Theory of international trade: a dual, general equilibrium approach*, Cambridge University Press.

Grewal, D., Levy, M., & Kumar, V. (2009). Customer experience management in retailing: An organizing framework. *Journal of Retailing*, 85(1), 1-14.

Guy, F. (2013). Small, local and cheap? Walkable and car-oriented retail in competition. *Spatial Economic Analysis*, 8(4), 425-442.

Helpman, E. (1984). Increasing returns, imperfect markets, and trade theory. *Handbook of international economics*, 1, 325-365.

Helpman, E., & Krugman, P. R. (1985). *Market structure and foreign trade: increasing returns, imperfect competition, and the international economy*, MIT press.

Huff, D. L. (1962). A note on the limitation of interurban gravity models. *Land Economics*, 38(1), 64-66.

Huff, D. L. (1963). Probabilistic analysis of shopping centre trade areas. *Land Economics*, 39, 81-90.

Huff, D. L. (2003). Parameter estimation in the Huff model, ESRI, *ArcUser*, 34-36.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *The American Economic Review*, 70(5), 950-959.

Kwan, M. P., & Weber, J. (2008). Scale and accessibility: Implications for the analysis of land use-travel interaction. *Applied Geography*, 28(2), 110-123.

Li, Y., & Liu, L. (2012). Assessing the impact of retail location on store performance: a comparison of Wal-Mart and Kmart stores in Cincinnati. *Applied Geography*, 32(2), 591-600.

Markusen, J. R. (1981). Trade and the gains from trade with imperfect competition. *Journal of international economics*, 11(4), 531-551.

Melitz, M. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity, *Econometrica*, 71, 1695-1725.

Melitz, M. & Ottaviano, G.I.P. (2008). Market size, trade, and productivity, *Review of Economic Studies*, 75, 295-316.

Nakanishi, M., & Cooper, L. G. (1974). Parameter estimation for a multiplicative competitive interaction model: least squares approach. *Journal of Marketing Research*, 303-311.

Nakanishi, M., & Cooper L.G. (1982). Simplified estimation procedures for MCI models, *Marketing Science*, 1(3), 314-322.

Olley, S. & Pakes, A. (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, 64(6), 1263-1297.

Öner, Ö (2017). Retail city: the relationship between place attractiveness and accessibility to shops. *Spatial Economic Analysis*, 12(1), 72-91.

Pavcnik, N. (2002). Trade liberalization, exit, and productivity improvements: evidence from Chilean plants, *Review of Economic Studies*, 69, 245-276.

Reilly, W. J. (1929). Methods for the study of retail relationships. *University of Texas Bulletin*, 2944.

Reilly, W. J. (1931). *The law of retail gravitation*. New York: Knicker-bocker Press.

Satani, N., Uchida, A., Deguchi, A., Ohgai, A., Sato, S., & Hagishima, S. (1998). Commercial facility location model using multiple regressions analysis. *Computer, Environment, and Urban Systems*, 22(3), 219-240.

Schivardi, F., & Viviano, E. (2011). Entry barriers in retail trade. *The Economic Journal*, 121(551), 145-170.

Suàrez-Vega, R., Santos-Peñate, D., Dorta-Gonzàlez, P., & Rodrìguez-Dìaz, M. (2011). A multi-criteria GIS based procedure to solve a network competitive location problem. *Applied Geography*, 31, 282e291.

Van Beveren, I. (2012). Total factor productivity estimation: a practical review. *Journal of Economic Surveys*, **26**(1), 98-128.

Wegener, M., Eskelinnen, H., Fürst, F., Schürmann, C., Spiekermann, K. (2002). *Criteria for the Spatial Differentiation of the EU Territory: Geographical Position*. Forschungen 102.2, Bonn: Bundesamt für Bauwesen und Raumordnung.

Yasar, M., Raciborski, R., & Poi, B. (2008). Production function estimation in Stata using the Olley and Pakes method. *Stata Journal*, 8(2), 221.

Table 1: Benchmark results.

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	AVG	AVG	AVG	SD	SD	SD
AMUSA	-0.105***	-0.061*		-0.245***	-0.186***	
	(0.03)	(0.04)		(0.07)	(0.07)	
Rental price of the surface area	0.034	0.181***	0.152**	-0.204*	-0.081	-0.130
	(0.06)	(0.06)	(0.06)	(0.11)	(0.11)	(0.11)
North dummy (λ_3^N)		-0.121***	0.106		-0.249***	-0.218
		(0.04)	(0.33)		(0.06)	(0.56)
South dummy (λ_3^S)		0.101***	0.938***		0.013	1.438***
		(0.04)	(0.29)		(0.07)	(0.53)
AMUSA*North (λ_4^N)			-0.062			-0.010
			(0.09)			(0.15)
AMUSA*South (λ_4^S)			-0.239***			-0.416***
			(0.08)			(0.15)
AMUSA - Center (λ_1)			0.062			-0.014
			(0.07)			(0.11)
Constant (λ_0)	-2.705***	-3.129***	-3.527***	-2.375***	-2.722***	-3.246***
	(0.14)	(0.16)	(0.24)	(0.26)	(0.30)	(0.37)
AMUSA - North ($\lambda_1 + \lambda_4^N$)			0.001			-0.024
			(0.06)			(0.11)
AMUSA - South ($\lambda_1 + \lambda_4^S$)			-0.177***			-0.430***
			(0.05)			(0.11)
N	3656	3656	3656	1982	1982	1982
N (North)	1790	1790	1790	901	901	901
N (Center)	630	630	630	375	375	375
N (South)	1236	1236	1236	706	706	706
adj R^2	0.002	0.013	0.014	0.015	0.026	0.029
p-value $H_0 : \lambda_1 <= 0$	.999	0.960	0.179	1.000	0.996	0.552
p-value $H_0 : \lambda_4^N <= 0$			0.754			0.526
p-value $H_0 : \lambda_4^S <= 0$			0.997			0.997
p-value $H_0 : (\lambda_1 + \lambda_4^N) <= 0$			0.495			0.584
p-value $H_0 : (\lambda_1 + \lambda_4^S) <= 0$			1.000			1.000

Standard errors are in parentheses. All continuous variables are in logs.
Two-sided test levels of significativeness: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Alternative specifications.

Dep. Var.:	(1) AVG	(2) AVG	(3) AVG	(4) SD	(5) SD	(6) SD
AMUSA-province	-0.045***			0.050*		
	(0.02)			(0.03)		
Regional Accessibility		-0.014			0.056*	
		(0.02)			(0.03)	
PAFS			-0.073***			-0.085***
			(0.02)			(0.03)
Financial Risk			-0.043			0.162*
			(0.05)			(0.10)
Rental price of the surface area	0.137**	0.151**	0.158**	-0.109	-0.114	0.026
	(0.06)	(0.06)	(0.06)	(0.11)	(0.11)	(0.13)
North dummy (λ_3^N)	0.162	0.078	-0.389	-0.325	-0.134	0.080
	(0.33)	(0.32)	(0.35)	(0.56)	(0.57)	(0.65)
South dummy (λ_3^S)	1.008***	0.933***	1.201***	1.295**	1.420***	1.669***
	(0.30)	(0.29)	(0.30)	(0.54)	(0.53)	(0.58)
AMUSA*North (λ_4^N)	-0.094	-0.060	0.073	0.039	-0.008	-0.085
	(0.09)	(0.09)	(0.09)	(0.15)	(0.15)	(0.18)
AMUSA*South (λ_4^S)	-0.269***	-0.241***	-0.306***	-0.361**	-0.397***	-0.455***
	(0.09)	(0.08)	(0.08)	(0.16)	(0.15)	(0.16)
AMUSA - Center (λ_1)	0.086	0.061	0.112*	-0.046	-0.007	0.016
	(0.07)	(0.07)	(0.07)	(0.11)	(0.11)	(0.12)
Constant (λ_0)	-3.659***	-3.568***	-3.903***	-3.092***	-3.120***	-4.266***
	(0.25)	(0.24)	(0.27)	(0.38)	(0.38)	(0.47)
AMUSA - North ($\lambda_1 + \lambda_4^N$)	-0.008	-0.006	0.000	-0.015	0.185*	-0.069
	(0.06)	(0.11)	(0.06)	(0.11)	(0.07)	(0.13)
AMUSA - South ($\lambda_1 + \lambda_4^S$)	-0.183***	-0.407***	-0.180***	-0.403***	-0.194***	-0.439***
	(0.05)	(0.11)	(0.05)	(0.11)	(0.05)	(0.11)
N	3656	3656	2991	1982	1982	1639
N (North)	1790	1790	1265	901	901	630
N (Center)	630	630	490	375	375	303
N (South)	1236	1236	1236	706	706	706
adj R^2	0.017	0.014	0.027	0.031	0.031	0.049
p-value $H_0 : \lambda_1 <= 0$	0.108	0.186	0.048	0.661	0.525	0.449
p-value $H_0 : \lambda_4^N <= 0$	0.845	0.749	0.219	0.398	0.522	0.683
p-value $H_0 : \lambda_4^S <= 0$	0.999	0.998	1.000	0.989	0.995	0.997
p-value $H_0 : (\lambda_1 + \lambda_4^N) <= 0$	0.551	0.522	0.497	0.553	0.003	0.070
p-value $H_0 : (\lambda_1 + \lambda_4^S) <= 0$	1.000	1.000	1.000	1.000	1.000	1.000

Standard errors in parentheses. All continuous variables are in logs.
Two-sided test levels of significativeness: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

A Huff model with heterogeneous retailers. Application to Italy

ONLINE APPENDIX*

This version: March 6, 2018

Massimo Del Gatto[1]
'G.d'Annunzio' University and CRENoS

Carlo S. Mastinu
University of Cagliari and CRENoS

*

[1]Corresponding author. Contact details: Massimo Del Gatto; Università 'G.d'Annunzio', Department of Economics (DEc); Viale Pindaro, 42; 65127 - PESCARA (ITALY); massimo.delgatto@gmail.com.

A Appendix: Descriptive statistics

Table ?? reports descriptive statistics for the main variables used in the analysis.[2]

Figure ?? visualizes the geographic distribution of the stores included in our dataset. The computed values of Ψ_f (see Section 4 in the paper) are reported in Figure ??.

B Appendix: Robustness checks

In this Appendix we report the robustness analysis for the results discussed in Section 5.

As a first experiment, we use alternative measures of central tendency and dispersion. In fact, the mean is not a good measure of central tendency in skewed distributions (like the Pareto for $k > 1$). For such distributions, the median is a better measure, and in our model it is equal to $\text{MDN}(c_f) = (0.5)^{\frac{1}{k}}\,\bar{c}_f$. For dispersion, a better measure in skewed distributions, compared to the standard deviation, can be the "interquartile range", defined as the difference between the 75th and the 25th percentiles. In our model, this is equal to $\text{IQR}(c_f) = \left[(0.75)^{\frac{1}{k}} - (0.25)^{\frac{1}{k}}\right]\bar{c}_f$. Both these measures are increasing in the UIR cutoff. The results of the alternative regressions including the median and the IQR are reported in columns 1-3 and 4-6 of Table ??, respectively. The benchmark results are confirmed.

In a second robustness check, we recognize that municipalities with highest accessibility are Rome and Milan, located in Central and Northern Italy, respectively. If the selection effect is particularly low in these two provinces, we are likely to estimate a not significant AMUSA effect for the Central and Northern municipalities. The regressions in Table ??, run without the municipalities located in the provinces of Rome and Milan, show that this is not the case.

Our dataset covers four categories of stores (hypermarkets, supermarkets, discounters, and small retailers). In principle, the results might differ across categories. In particular, one might think of the market potential of supermarkets, discounters, and small retailers as associated to a smaller geographical scale. In Table ?? we show that the benchmark results remain valid when only hypermarkets are considered.

The AMUSA is a composition of distance-weighted consumers' income and distance-weighted store size (i.e., Ψ_f and Θ_A respectively). In principle, one would expect these two terms to be in a negative and positive relationship with the UIR average and dispersion, respectively. In Table ??, we consider the two variables separately and show

[2] Notice that the Table reports the UIC values before taking logs. The presence of negative values is due to the presence of municipalities with a prevalence of stores characterized by a negative value added. These observations do not take part in the estimation, which is carried out in logarithms. This explains the different number of observations with respect to the regression output tables.

that this is indeed the case.

Finally, we address the question: why not going deeper in the geographical disaggregation and estimate region-specific AMUSA effects and UIR distributions? A first answer is that we are concerned with providing 'general results' concerning the effectiveness of the selection effect. By this perspective, our analysis requires detailed data, on the one hand, as the reference unit is the retail trade area, and a wide geographical scale, on the other hand, in order to avoid results that are too much specific. Nevertheless, we carried out regressions (analogous to the ones displayed in Table 1 of the paper) at the regional level. The output of this exercise, reported in Table **??**, shows that switching to the regional scale highly blurs the analysis of the selection effect. In many cases, the AMUSA effect even becomes positive (note that the excluded region is always Region 9: Lazio). This notwithstanding, since the AMUSA in Table **??** is negative and significative in all Southern regions but Calabria and Campania, the analysis can be seen as a sort of regional disaggregation of the AMUSA effect documented for Southern Italy in our benchmark regressions.

Two points are worth noting. First, as confirmed by Tables **??** and **??**, where the AVG and SD regressions are run region by region, the number of observations at the regional level becomes sometimes quite low and is likely to lead to inconsistent estimates in Table **??**. This is particularly true in the dispersion analysis, where the number of observations shrinks even more considerably due to the presence of municipalities including only one store. Second, differently from what could be inferred from Table **??**, the by-region analysis in Tables **??** and **??** reveals an insignificant AMUSA effect in almost all regions. This is not the case in Table **??**, where the benchmark regressions are carried out again by macro-region: in this case, the output is fully consistent with the benchmark results, entailing that the results in Table **??** are mostly driven by the the rental price of the surface area. All in all, these arguments suggest that, although interesting in principle, the regional tables presented in this section have to be taken with caution, since our dataset is not best suited for addressing the regional scale of the selection process.

Table A1: Descriptive statistics: main variables by region.

Region	$AVG(c)$	$MDN(c)$	$Min(c)$	$Max(c)$	$SD(c)$	$IQR(c)$	AMUSA	Ψ	Θ_A	ω	N of mun.
EMILIA-ROMAGNA	0.04912	0.04820	0.00067	0.11218	0.02975	0.02836	59.88	936.80	17.05	7.99	221
FRIULI-VENEZIA G	0.05416	0.05411	0.01378	0.13551	0.02761	0.02348	36.46	686.92	21.97	6.83	110
LIGURIA	0.04508	0.04552	0.00067	0.10235	0.03019	0.02064	59.99	421.04	8.15	9.25	64
LOMBARDIA	0.05528	0.05531	-0.58982	0.87920	0.03651	0.02355	53.72	2262.51	42.94	8.22	635
PIEMONTE	0.01393	0.00910	-12.93994	1.72056	0.09392	0.05196	66.41	689.02	16.32	6.26	198
TRENTINO-ALTO AD	0.07896	0.08009	0.00067	0.12248	0.02241	0.01251	54.89	225.09	4.98	9.87	168
VALLE D'AOSTA	0.03756	0.03756	0.00086	0.08639	0.00331	0.00156	61.62	148.23	3.29	8.13	6
VENETO	0.04666	0.04757	-0.30128	0.16463	0.02578	0.02525	42.21	1057.92	29.20	7.99	401
N	0.04979	0.04947	-3.67860	0.80820	0.03616	0.02643	55.55	1233.24	25.88	7.76	1803
LAZIO	0.02414	0.01577	-6.32947	0.87341	0.04601	0.04061	49.58	470.54	11.14	8.38	207
MARCHE	0.05722	0.05724	0.00079	0.11683	0.03494	0.02842	40.37	342.19	10.66	7.07	141
TOSCANA	0.07516	0.07541	0.00067	0.16360	0.03381	0.03240	63.41	713.53	11.88	10.12	210
UMBRIA	0.05821	0.05490	0.02425	0.11683	0.03296	0.03833	36.57	316.03	9.77	6.21	76
C	0.05248	0.04944	-2.39403	0.41678	0.03775	0.03491	50.17	495.32	11.11	8.37	634
ABRUZZO	0.03738	0.03640	0.00067	0.09991	0.02533	0.02071	58.77	204.83	6.50	6.14	109
BASILICATA	0.06108	0.05953	0.01375	0.64408	0.07052	0.05385	44.68	76.84	2.48	4.87	44
CALABRIA	0.14121	0.11806	-0.20146	1.85520	0.14493	0.08531	33.99	100.26	4.80	4.49	152
CAMPANIA	0.05438	0.06003	-1.64187	0.63675	0.04648	0.02410	44.54	908.76	24.22	5.69	304
MOLISE	0.04335	0.04483	0.00105	0.07443	0.02472	0.01979	83.94	89.58	2.25	4.96	28
PUGLIA	0.01781	0.01742	-5.61284	0.39588	0.05167	0.03347	43.24	597.55	19.38	6.06	200
SARDEGNA	0.08923	0.07633	-0.28045	1.47395	0.08069	0.03554	30.79	156.90	6.61	8.12	236
SICILIA	0.03615	0.01818	-7.94656	1.21609	0.10191	0.04801	34.75	387.54	12.61	5.73	222
S-I	0.06071	0.05372	-2.20577	0.92568	0.07019	0.03955	42.65	387.65	11.87	5.85	1295
Total	0.05403	0.05094	-3.05520	0.79727	0.04906	0.03242	50.81	875.28	19.63	7.24	3732

Table B1: Robustness: UIR median and IQR.

	(1) $MDN(c)$	(2) $MDN(c)$	(3) $MDN(c)$	(4) $IQR(c)$	(5) $IQR(c)$	(6) $IQR(c)$
AMUSA (Φ_m)	-0.068**	-0.034	0.135*	-0.187***	-0.170**	0.090
	(0.03)	(0.04)	(0.07)	(0.07)	(0.07)	(0.12)
Rental price of land (ω_p)	0.016	0.137**	0.111*	-0.130	-0.125	-0.186
	(0.06)	(0.06)	(0.06)	(0.12)	(0.13)	(0.13)
North (dummy)		-0.069*	0.562*		-0.233***	0.186
		(0.04)	(0.34)		(0.07)	(0.62)
South (dummy)		0.104**	0.958***		-0.133*	1.704***
		(0.04)	(0.30)		(0.08)	(0.57)
AMUSA*North			-0.170*			-0.116
			(0.09)			(0.17)
AMUSA*South			-0.239***			-0.533***
			(0.08)			(0.16)
Constant	-2.856***	-3.222***	-3.794***	-2.460***	-2.377***	-3.197***
	(0.14)	(0.17)	(0.25)	(0.26)	(0.32)	(0.43)
N	3667	3667	3667	1960	1960	1960
adj R^2	0.001	0.006	0.007	0.007	0.012	0.017

Standard errors are in parentheses. All continuous variables are in logs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B2: Robustness: excluding Rome and Milan (highest accessibility).

	(1) $AVG(c)$	(2) $AVG(c)$	(3) $AVG(c)$	(4) $SD(c)$	(5) $SD(c)$	(6) $SD(c)$
AMUSA (Φ_m)	-0.105***	-0.061*	0.064	-0.246***	-0.187***	-0.016
	(0.03)	(0.04)	(0.07)	(0.07)	(0.07)	(0.11)
Rental price of land (ω_p)	0.034	0.182***	0.153**	-0.207*	-0.084	-0.132
	(0.06)	(0.06)	(0.06)	(0.11)	(0.11)	(0.11)
North (dummy)		-0.122***	0.112		-0.249***	-0.216
		(0.04)	(0.33)		(0.06)	(0.56)
South (dummy)		0.100**	0.943***		0.012	1.433***
		(0.04)	(0.29)		(0.07)	(0.53)
AMUSA*North			-0.063			-0.011
			(0.09)			(0.15)
AMUSA*South			-0.241***			-0.415***
			(0.08)			(0.15)
Constant	-2.706***	-3.130***	-3.533***	-2.366***	-2.713***	-3.238***
	(0.14)	(0.16)	(0.24)	(0.26)	(0.30)	(0.37)
N	3654	3654	3654	1980	1980	1980
adj R^2	0.002	0.013	0.015	0.015	0.026	0.029

Standard errors are in parentheses. All continuous variables are in logs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B3: Robustness: hypermarkets only.

	(1) $AVG(c)$	(2) $AVG(c)$	(3) $AVG(c)$	(4) $SD(c)$	(5) $SD(c)$	(6) $SD(c)$
AMUSA (Φ_m)	-0.105***	-0.061*	0.062	-0.245***	-0.186***	-0.014
	(0.03)	(0.04)	(0.07)	(0.07)	(0.07)	(0.11)
Rental price of land (ω_p)	0.034	0.181***	0.152**	-0.204*	-0.081	-0.130
	(0.06)	(0.06)	(0.06)	(0.11)	(0.11)	(0.11)
North (dummy)		-0.121***	0.106		-0.249***	-0.218
		(0.04)	(0.33)		(0.06)	(0.56)
South (dummy)		0.101***	0.938***		0.013	1.438***
		(0.04)	(0.29)		(0.07)	(0.53)
AMUSA*North			-0.062			-0.010
			(0.09)			(0.15)
AMUSA*South			-0.239***			-0.416***
			(0.08)			(0.15)
Constant	-2.705***	-3.129***	-3.527***	-2.375***	-2.722***	-3.246***
	(0.14)	(0.16)	(0.24)	(0.26)	(0.30)	(0.37)
N	3656	3656	3656	1982	1982	1982
adj R^2	0.002	0.013	0.014	0.015	0.026	0.029

Standard errors are in parentheses. All continuous variables are in logs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B4: Robustness: regressions with AMUSA decomposed into accessibility and competition.

	(1) $AVG(c)$	(2) $AVG(c)$	(3) $SD(c)$	(4) $SD(c)$
Accessibility (Ψ_m)	-0.123***	-0.078**	-0.240***	-0.188***
	(0.03)	(0.04)	(0.07)	(0.07)
Total Sales Area (Θ_A)	0.082**	0.052	0.223***	0.199***
	(0.03)	(0.04)	(0.07)	(0.07)
Rental price of land (ω_p)	0.076	0.192***	-0.200*	-0.082
	(0.06)	(0.06)	(0.11)	(0.11)
North (dummy)		-0.095**		-0.261***
		(0.04)		(0.06)
South (dummy)		0.098**		0.009
		(0.04)		(0.07)
Constant	-2.627***	-3.040***	-2.357***	-2.737***
	(0.14)	(0.17)	(0.27)	(0.30)
N	3656	3656	1982	1982
adj R^2	0.008	0.014	0.014	0.025

Standard errors are in parentheses. All continuous variables are in logs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B5: Regressions by macro-region.

	$AVG(c)$ North	$AVG(c)$ Center	$AVG(c)$ South	$SD(c)$ North	$SD(c)$ Center	$SD(c)$ South
AMUSA	0.021	0.049	-0.181***	-0.033	-0.013	-0.428***
Rental price of the surface area	0.004	0.202*	0.283***	-0.079	-0.134	-0.182
Constant	-3.189***	-3.582***	-2.808***	-3.537***	-3.243***	-1.726***
N	1790	630	1236	901	375	706
adj R^2	0	0	0.02	0	0	0.02

Standard errors are in parentheses. All continuous variables are in logs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B6: Regional regressions.

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	$AVG(c)$	$AVG(c)$	$AVG(c)$	$SD(c)$	$SD(c)$	$SD(c)$
AMUSA (Region 9)	-0.105	0.035	-0.015	-0.245	-0.085*	-0.456**
Rental price of the surface area	0.034	-0.035	-0.025	-0.204	0.182	0.195
Dummy Region 1		-0.278***	2.307***		0.150***	0.419*
Dummy Region 2		0.080**	0.250**		-0.198**	0.984**
Dummy Region 3		-0.479***	-3.226***		-0.056***	-5.185***
Dummy Region 4		0.011	-1.308***		-0.267***	-1.739***
Dummy Region 5		0.154***	0.435***		0.557***	-2.476***
Dummy Region 6		0.479***	1.376***		-0.037*	-0.118
Dummy Region 7		-1.161***	-14.497***		-2.906***	-15.932***
Dummy Region 8		-0.042***	-0.526***		-0.145***	-0.749**
Dummy Region 10		0.03	2.255***		0.343***	-1.479***
Dummy Region 11		0.319***	0.107**		0.163***	-2.349***
Dummy Region 12		0.169***	-0.263**		0.308***	0.723**
Dummy Region 13		-0.339***	-0.064*		-0.041	-1.292**
Dummy Region 14		-0.009	0.228***		0.279**	-2.756***
Dummy Region 15		0.192*	0.082***		0.991***	0.884**
Dummy Region 16		0.128**	-0.186***		0.069	-1.237***
Dummy Region 17		-0.252***	2.426***		-0.032	4.335***
Dummy Region 18		-0.022	0.016		0.042	-3.000***
Dummy Region 19		0.559***	0.624***		0.350***	-0.923**
Dummy Region 20		0.556***	0.635***		0.817***	0.007
AMUSA * Dummy Region 1			-0.662***			-0.492**
AMUSA * Dummy Region 2			-0.067***			-0.828***
AMUSA * Dummy Region 3			0.690***			0.915***
AMUSA * Dummy Region 4			0.334***			-0.047
AMUSA * Dummy Region 5			-0.093**			0.390*
AMUSA * Dummy Region 6			-0.253**			-0.428*
AMUSA * Dummy Region 7			3.674			3.150***
AMUSA * Dummy Region 8			0.119***			-0.288***
AMUSA * Dummy Region 10			-0.657***			0.060**
AMUSA * Dummy Region 11			0.042**			0.211**
AMUSA * Dummy Region 12			0.107***			-0.592***
AMUSA * Dummy Region 13			-0.093***			-0.106***
AMUSA * Dummy Region 14			-0.087***			0.441***
AMUSA * Dummy Region 15			0.013**			-0.500***
AMUSA * Dummy Region 16			0.073**			-0.090*
AMUSA * Dummy Region 17			-0.724***			-1.677***
AMUSA * Dummy Region 18			-0.029*			0.423**
AMUSA * Dummy Region 19			-0.043***			-0.102*
AMUSA * Dummy Region 20			-0.043***			-0.242***
Constant	-2.705**	-3.167***	-3.006***	-2.375*	-3.814***	-2.516***
N	3656	3656	3656	1982	1982	1982
adj R^2	0	0.08	0.09	0.01	0.08	0.09

Standard errors clustered by macro-region (North, Center, South) are in parentheses. All continuous variables are in logs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Region 1 =”EMILIA-ROMAGNA”, Region 2=”FRIULI-VENEZIA, Region 3 =”LIGURIA”
Region 4 =”LOMBARDIA”, Region 5 =”PIEMONTE”, Region 6 =”TRENTINO-ALTO
Region 7 =”VALLE D’AOSTA”, Region 8 =”VENETO”, Region 9 =”LAZIO”
Region 10 =”MARCHE”, Region 11 =”TOSCANA”, Region 12 =”UMBRIA”
Region 13 =”ABRUZZO”, Region 14 =”BASILICATA”, Region 15 =”CALABRIA”
Region 16 =”CAMPANIA”, Region 17 =”MOLISE”, Region 18 =”PUGLIA”
Region 19 =”SARDEGNA”, Region 20 =”SICILIA”

Table B7: Regressions by region. Dependent variable: $AVG(c)$.

	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6	Region 7	Region 8	Region 9	Region 10
AMUSA	-0.364	-0.061	0.666	0.322***	-0.287	-0.139	3.674**	0.108	0.014	-0.880**
Rental price of the surface area	-2.241***	-1.134	-0.271	0.077	1.251**	-0.915**	(omitted)	0.267*	-0.19	4.611***
Constant	2.722	-0.653	-5.591	-4.485***	-4.244***	-0.07	-17.554**	-4.093***	-2.765***	-9.024***
N	221	110	64	628	194	168	6	399	204	140
adj R^2	0.06	0	0.03	0.02	0.03	0.07	0.19	0	0	0.12

	Region 11	Region 12	Region 13	Region 14	Region 15	Region 16	Region 17	Region 18	Region 19	Region 20
AMUSA	-0.033	0.081	-0.087	-0.079	-0.043	0.083	-0.757	0.006	-0.045	-0.046
Rental price of the surface area	0.571***	-0.64	-1.316	-2.467	-4.447***	-0.092	-2.334	-0.291	0.230*	-0.65
Constant	-3.970***	-2.052	-0.749	1.067	3.891*	-3.109***	3.272	-2.629***	-2.902***	-1.268
N	210	76	107	44	150	303	28	197	217	190
adj R^2	0.01	-0.01	0	-0.04	0.06	0	0.11	-0.01	0	0

Standard errors are in parentheses. All continuous variables are in logs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B8: Regressions by region. Dependent variable: $SD(c)$.

	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6	Region 7	Region 8	Region 9	Region 10
AMUSA	-0.453	-0.867*	1.391	-0.098	0.236	0.156	3.15	-0.285	-0.539	0.003
Rental price of the surface area	-0.073	2.371	3.891	0.486	0.82	-1.715	(omitted)	0.118	0.402	1.767
Constant	-1.689	-5.555	-17.679*	-4.686***	-5.606***	-0.437	-18.039	-3.116***	-2.663*	-6.870***
N	119	61	28	291	91	50	2	259	120	72
adj R^2	0	0.02	0.13	0	0.01	0.08	.	0	0	0

	Region 11	Region 12	Region 13	Region 14	Region 15	Region 16	Region 17	Region 18	Region 19	Region 20
AMUSA	0.259	-0.619	-0.115	0.442	-0.596**	-0.163	-1.576	0.435	-0.099	-0.223
Rental price of the surface area	-0.705**	-0.89	-0.475	0.078	-5.296**	0.543	-2.038	0.143	0.24	-1.940*
Constant	-2.978***	0.291	-2.566	-5.09	6.861*	-4.111***	5.018	-5.462***	-3.543***	1.167
N	126	57	58	23	64	189	15	148	97	112
adj R^2	0.01	0.05	-0.03	-0.08	0.08	0.01	0.24	0	-0.02	0.01

Standard errors are in parentheses. All continuous variables are in logs.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Region 1 =”EMILIA-ROMAGNA”, Region 2=”FRIULI-VENEZIA, Region 3 =”LIGURIA”
Region 4 =”LOMBARDIA”, Region 5 =”PIEMONTE”, Region 6 =”TRENTINO-ALTO
Region 7 =”VALLE D’AOSTA”, Region 8 =”VENETO”, Region 9 =”LAZIO”
Region 10 =”MARCHE”, Region 11 =”TOSCANA”, Region 12 =”UMBRIA”
Region 13 =”ABRUZZO”, Region 14 =”BASILICATA”, Region 15 =”CALABRIA”
Region 16 =”CAMPANIA”, Region 17 =”MOLISE”, Region 18 =”PUGLIA”
Region 19 =”SARDEGNA”, Region 20 =”SICILIA”

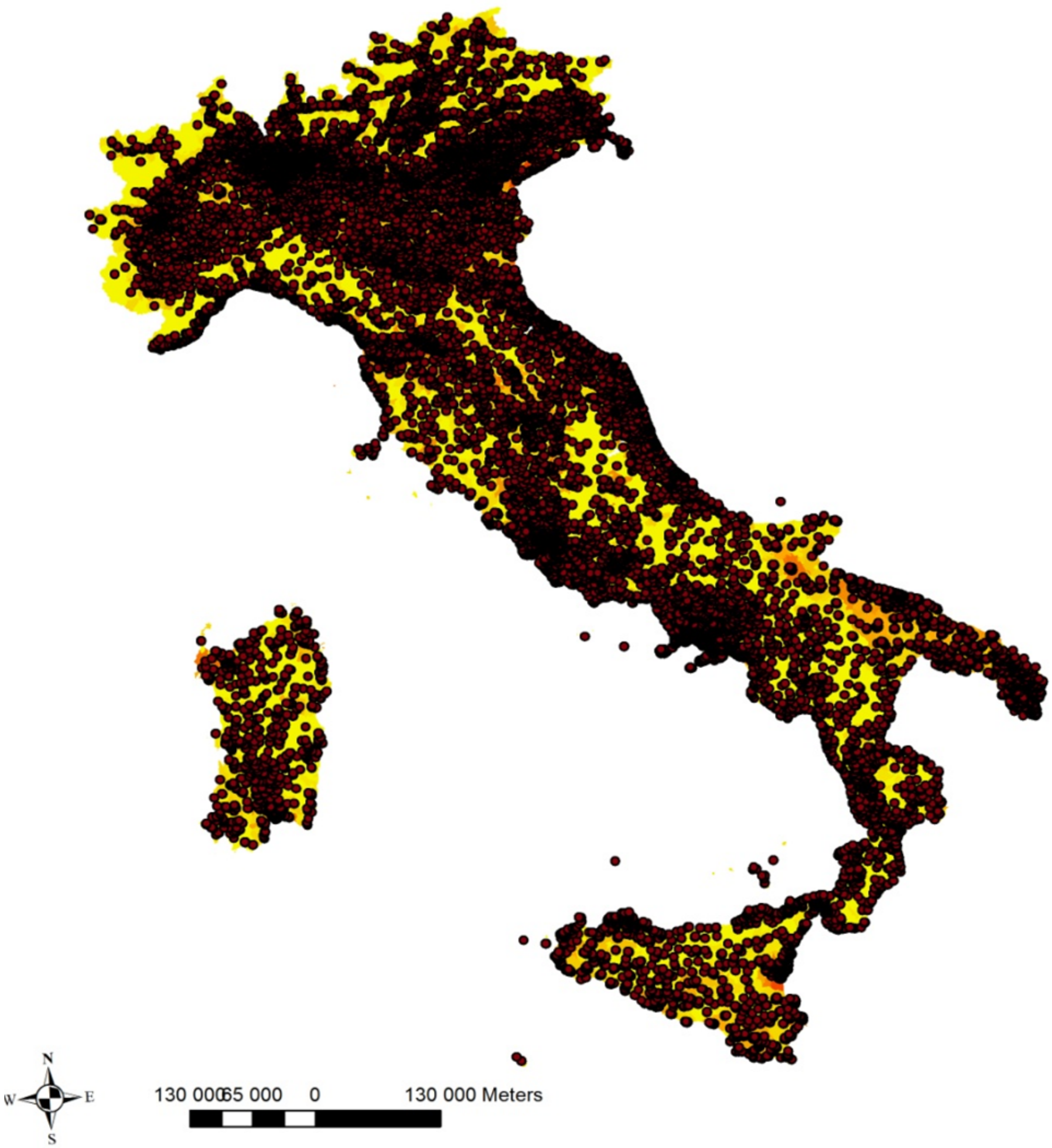

Figure A1: Geographic distribution of the stores in the Nielsen database.

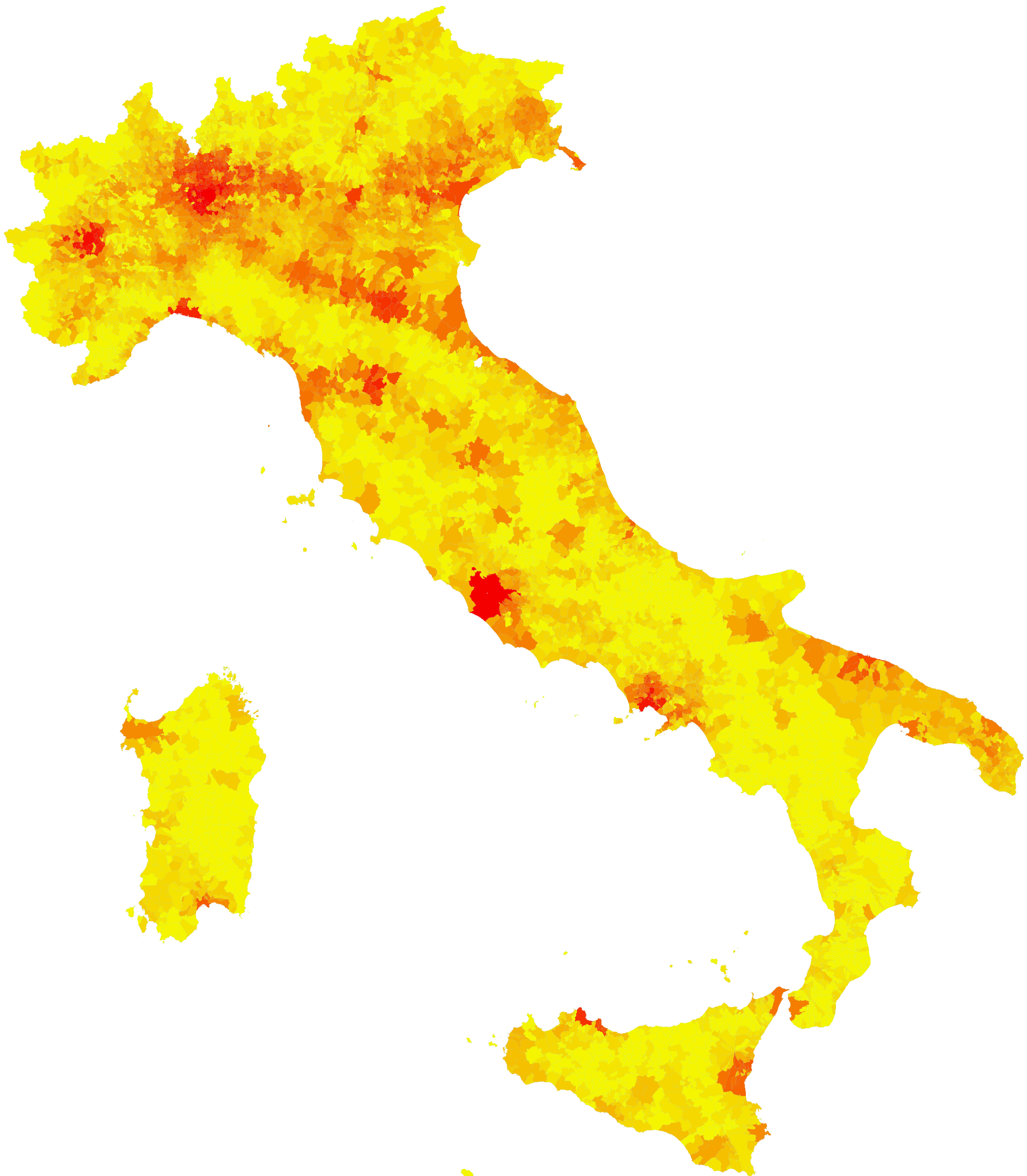

Figure A2: Potential accessibility at the municipality level (Ψ_f).