Towards an Integrated Methodology for Model and Variable Selection Using Count Data: An Application to Micro-Retail Distribution in Urban Studies

Alessandro Araldi

urban science

MDPI

Article

Towards an Integrated Methodology for Model and Variable Selection Using Count Data: An Application to Micro-Retail Distribution in Urban Studies

Alessandro Araldi

ESPACE, CNRS, University Côte d'Azur, 06200 Nice, France; alessandro.araldi@univ-cotedazur.fr

Received: 23 March 2020; Accepted: 26 April 2020; Published: 28 April 2020

check for updates

Abstract: Over the last two decades, a growing number of works in urban studies have revealed how micro-retail distribution is significantly related to specific properties of the urban built environment. While a wide variety of urban form measures have been investigated using sophisticated analytical approaches, the same attention has not equally been found in statistical procedures. Several essential features of micro-retail statistical distribution and modelling assumptions are frequently overlooked, compromising the statistical robustness of outcomes. In this work we focus on four main aspects: (i) the discrete, non-negative and highly skewed nature of store distribution; (ii) its zero-inflation; (iii) assessment of the contextual effect; and (iv) the multicollinearity generated by the inclusion of highly related urban descriptors. To overcome these limitations, we propose an integrated methodological framework for both modelling and variable selection assessment based on generalized linear models (GLMs) and elastic-net (Enet) penalized regression (PR), respectively. The procedure is tested via a real case study of the French Riviera, which is described using a large dataset of 105 street-based urban form measures. The outcomes of this procedure show the superiority of the zero-inflate negative binomial count regression approach. A restricted number of urban form properties are found to be related to the micro-retail distribution depending on the specific scale and morphological context under analysis.

Keywords: micro-retail distribution; street-based urban form; skeletal streetscape; street-network configuration; Penalized Regression; zero-inflated negative binomial regression; Variable selection

1. Introduction

Stores represent one of the most important elements of the urban environment. Their presence engenders human interaction, socioeconomic vibrancy, cohesion and sense of place from the street to the neighborhood level, ultimately affecting the attractiveness of a whole city [1,2]. In the last two decades, the increasing availability of micro-store data has stimulated a growing body of research investigating the different location factors underlying store distribution [3]. More particularly, quantitative urban geography and urban form studies have explored the relationship between micro-retail distribution and the physical properties of urban form, also named "the morphological sense of commerce" [4]. The goal underlying these works is to investigate if and how the spatial organizations of urban form elements (streets, building and plots) influence the human perception and usages of urban spaces, and, subsequently, whether this effects the distribution of socio-economic activities such as traditional brick and mortar micro-retailers. These works might provide academics and practitioners with evidence on how urban systems work and nourish discussion about how to improve life quality in urban areas through their design and planning.

Among the forerunners of this specific research stream, we find Hillier's Movement Economy Theory (MET) [5–7]. MET explains how the spatial configuration of public spaces influences

movement patterns and, indirectly, the localization of stores. Several protocols and modelling approaches have highlighted the importance of different street-network configurational properties on micro-retail distribution. Micro-retail patterns have proven to be significantly correlated with integration (to-movement) and betweenness (through-movement) centrality measures defined within the Space Syntax-SSx [7] and Multiple Centrality Assessment (MCA) [8,9] methodological frameworks, respectively.

Although there is a general agreement about the importance of street-network properties, one main limitation concerns the absence of other essential features of urban morphology such as building distribution and height, site morphology, built-up density, and so on. Integrating these aspects might provide a more holistic description of urban form and, therefore, of its relationship with micro-retail distribution [10]. Together with configurational approaches, researchers have gradually introduced additional urban form descriptors evaluating both their individual and combined effects on micro-retail distribution, including street-based urban design qualities [11], street-block typologies and built-up density [12,13] and plot systems [14], among others. Moreover, other researchers have started investigating how the importance of each urban form variable might play different roles in micro-retail distribution depending on the relative morphological context defined as, for instance, city size, central/peripheral sectors [15] or underlying planned/spontaneous urban grids [16,17].

While sophisticated approaches have been developed for the identification, conceptualization and description of urban form, the same attention has not been equally found in modelling and statistical analysis. In this work, the implementation of well-established urban form analytical approaches provides us with a large number of street-based descriptors and allows us to focus on the statistical and modelling procedures implemented to describe the relationship between urban form and micro-retail distribution.

The first part of this paper is devoted to a critical review of the methodological procedures developed in previous works; the discussion is organized around four traditionally overlooked aspects:

(i) the discrete, non-negative and highly skewed nature of micro-retail distribution, which is incompatible with the assumptions underlying traditional statistical approaches;
(ii) the store absence, which characterizes urban spaces, is represented by a highly zero-inflated statistical distribution; store absence has both theoretically and methodologically been excluded from analysis;
(iii) the role of contextual descriptors and their inclusion in traditional regression approaches;
(iv) the presence of high multicollinearity when considering a large set of urban descriptors, which has been an issue in several methodological and theoretical approaches.

To overcome these limitations, an integrated methodological framework based on generalized linear models (GLMs) is herein proposed and implemented in order to study the relationship between micro-retail distribution and urban form in the French Riviera metropolitan area.

We begin by showing the superiority of the Zero-Inflated Negative Binomial (ZINB) model when applied to several regressive approaches. Beyond the ability of ZINB regressions to handle both skewness and an over-representation of zeroes, this outcome also supports the hypothesis of a double generative process describing two main aspects of micro-retail distribution in urban environments: presence/absence on one side, and the number/density of stores on the other. Moreover, the implementation of penalized regression (PR) as a built-in solution for variable selection procedures within the GLM framework allows the identification of specific subsets of urban morphological indicators depending on the urban typo-morphological context under investigation.

This work is part of a wider research project studying the relationship between urban form and the retail system, developing innovative methodologies for their study and producing new knowledge about the French Riviera urban system. This paper focuses on methodological developments and proposes an innovative procedure to model the relationship between retail distribution and urban form.

This paper is organized as follow: Section 2 provides a critical review of the statistical procedures developed in previous works, wherein the four aforementioned limitations are individually discussed. Section 3 describes the modelling and variable selection procedures based on GLM and PR, respectively. This protocol is then tested using a case study of the French Riviera, with outcomes discussed in Section 4. Limitations and future research perspectives conclude the paper.

2. Methodological Literature Review

2.1. Analysing the Relationship between Urban Form and Micro-retail Distribution

As presented above, several studies have explored the role played by different urban morphological aspects on micro-retail distribution. Table A1 presents a non-exhaustive collection of recent works investigating micro-retail locational factors related to the physical urban environment; for each paper, the column "analytical approach" highlights the statistical analytical/modelling approach implemented.

Three groups might be recognized: visual exploration of spatial co-occurrences, bivariate statistical tests and simple/multiple linear regressions (MLR).

Despite valuable observations highlighted in those works based on a visual exploration of spatial co-occurrences [13], weak and non-reproducible outcomes prevent both theoretical and methodological inferences or comparative analysis. Confirmation biases might also affect conclusions outlined through this approach.

Pearson's correlation represents the most implemented analytical procedure [8–10,18–20]. This approach allows the presence of a linear relationship between two continuous variables to be evaluated. Nonetheless, when considering micro-retail distribution, the normality assumption underlying Pearson's correlation test is not met: micro-retail measures can only assume positive values within the interval $[0, +\infty)$. Moreover, the presence of outliers and the high presence of zeroes (absence of stores) might further increase the distribution skewness. Similar characteristics can be associated with urban form descriptors (i.e., network centralities, building coverage ratios). Consequently, the statistical significance of correlation tests might result in biased outcomes.

Three main solutions might be considered: (i) implementation of a rank correlation test (i.e., Spearman, Kendall); (ii) evaluation of the non-linear fit between micro-retail and urban form properties [16,21]; and (iii) manipulations of the original data to meet a normality assumption such as log-transformation or, similarly, smoothing approaches [8–10,18–20]. Independently from the specific solution, these procedures propose simple bivariate analysis, assessing strength and direction of the relationship between each morphological variable and micro-retail measures. However, multivariate effects occur at the same time; stores might be detected through correspondence of a particular combination of variables. The simple bivariate correlation should not be implemented when the goal of the study (as in this work) is the evaluation of the combined importance of a set of explanatory variables.

MLR evaluates the combined effects of several independent descriptors, disentangling and examining their separate effects and assessing both partial and semi-partial correlations with target variables. To analyze the relationship between micro-retail and urban form, several works have implemented this modelling approach [17,22–25]. Although MLR is considered to be a powerful technique, it might not always represent the best solution. As previously discussed for statistical correlation tests, strong assumptions are also required for the implementation of MLR. The main assumption of homoscedasticity (normal distribution of residuals) is often not respected when a response variable is skewed (e.g., store count/density per street). In this case, residuals almost always correlate positively with the predictors and, consequently, the estimated standard errors of the regression coefficients are smaller than their true values. However, "heteroscedasticity is a problem with the model, not the data" [26].

Instead of discussing possible violations and subsequent data manipulations, the choice of a statistical modelling approach adapted to the nature of the data under analysis might be considered a simpler and much more effective solution. Although several retail distribution measures have been

used as response variables (i.e., store floor space, sales volume, workforce, frontage length) when studying the spatial distribution of stores, the simple number of establishments within a given region represents the most adopted solution. Individual stores represent the natural level of the analysis as well as the legal and functional unit for most businesses [27,28]. "Counting values represent the natural, obvious, and meaningful scale to describe discrete occurrences/distribution, and one should retain these virtues if possible" [26]. In this case, the dependent variable can only take discrete non-negative values, and it does not necessarily follow a normal distribution. In such cases, a conventional linear regression cannot be applied; instead of proposing ad-hoc transformations, count regression approaches should be preferred [29,30].

Count regression approaches have been developed since the end of the 1980s in different domains of study, including retail geography [27,31,32]. Nonetheless, among the academic community investigating urban form, only recent works have proposed the implementation of count regression approaches; for instance, Ye et al. [12] considered negative binomial regression when studying the relationship between street-block properties with catering-related stores. However, the superiority of this model to the more traditional MLR approach has not yet been evaluated. Therefore, the first goal of this work is to propose such an assessment for the modelling of the relationship between micro-retail distribution and urban form through the implementation of a robust model selection procedure.

2.2. *Stores Absence and the Survivorship Bias*

Beyond the choice of the most adapted analytical approach, a second aspect should be highlighted. The attention of academics is traditionally captured by those spaces where stores are observed, whereas absence is usually ignored or considered in the same way as missing data. Few works have tackled this specific feature. For instance, while correlating micro-retail presence and street-network configurational indicators, Omer and Goldblatt [17] compared values obtained from the overall study area to those observed in the subset of streets with a micro-retail presence different from zero.

The absence of stores in urban spaces has never been considered as an integral part of the process defining micro-retail spatial distribution, leading to what several disciplines have recognized as survivorship bias [33]. This bias is explained as the tendency of (statistic) studies to draw conclusions considering a subset of "successful" individuals who might not be representative of the overall population.

In our context, the emphasis on a specific subset of spatial units might be explained by the high heterogeneity characterizing micro-retail distribution in an urban space, which becomes even more evident when using fine-grained spatial units such as street segments. Micro-retail is found only in a small percentage of the total number of street-based units. Hence, academics traditionally apply a manual reduction of the zero overrepresentation [9,14,20,22]. Selecting and analyzing spatial units based on a specific criterion is not a bias in itself, although biases might arise in the interpretation of its results. This selection might be considered to be a legitimate choice under the condition that interpretations and conclusions of results highlight this background constraint underlying the statistical analysis. Thus, the same statistical relationships might or might not be verified when extending the analysis to the whole dataset. In other terms, outcomes explain the necessary but not sufficient conditions by which to observe the phenomena.

The exclusion of spatial units from the statistical analysis are not only a possible source of bias, but also represent a significant loss of knowledge. The absence of micro-retail should be considered just as predictive and informational as its presence. Integrating this aspect into the analysis would require limiting data manipulation procedures that lead to the manifestation of survivorship bias. Beyond the manual removal of zeros, smoothing/interpolation procedures such as kernel density estimation (KDE) [8–10,18–20] also result in similar conclusions. In this case, the relationship between the variables under analysis is overstated [24] (p. 63), valuable and detailed information such as an absence of stores might be omitted or diluted and the autocorrelation component of micro-retail distribution is artificially amplified.

Additionally, an important role is played by the spatial unit choice: precise and sharp information about store absence can be diluted depending on partition size (Modifiable Area Unit Problem-MAUP [34]). Two main strategies have been proposed when studying human-related phenomena: the use of behavioral-based scales [35] and reduction of the aggregation scale [36]. In this work, the use of street counts on street-based spatial units addresses both these requirements, allowing the investigation of both the presence and absence of stores.

Analyzing micro-retail absence also requires specific analytical/modelling procedures. Within the count regression methodological framework, explicit approaches have been proposed by which to evaluate zero-inflation in the target variable [37]. The hypothesis underlying these approaches is that two processes might generate the distribution of micro-retail activity: the first is responsible for its absence/presence, while the second explains its intensity. Therefore, we can investigate whether different combinations of urban morphological parameters underlie the two processes.

While zero-inflated procedures have been largely investigated in several domains, the appropriateness of these approaches still need to be assessed within the aforementioned urban form literature. Thus, the second goal of this work is to test the hypothesis regarding the presence of a double process that describes micro-retail presence/absence and magnitude through zero-inflated GLM approaches.

2.3. The Contextual Effect

So far we have discussed how specific modelling approaches should be considered when studying the relationship between micro-retail distribution and urban form. However, their statistical distribution and subsequent statistical relationship might vary depending on the specific urban context under analysis.

The contextual effect (also called the neighborhood, integral or landscape effect, depending on the discipline [38,39]) has previously been investigated only by a limited number of authors among the aforementioned studies. Among urban form literature, several works have integrated urban morphological context descriptors that have been defined using different approaches: expert-based knowledge incorporating urban and architectural data [16], official land-use zoning [24], center–periphery subsystem definitions [15], historical urban planning growth (planned/spontaneous) [17] and density types [14].

The spatial context represents a fundamental component of urban systems under study, especially when studying large regions encompassing heterogeneous urban forms. Overlooking this aspect might impact the model outcomes with a systematic over/underestimation effect and spurious correlations between dependent and independent variables [40,41]. Therefore, morphological-based partitions are included among the descriptor of the urban form in this work.

Several approaches have been developed to integrate contextual variables in statistical modelling. When referring to urban form literature, three main approaches have been considered. Scoppa [24] implemented both least square means and disaggregated approaches. Least square means is considered an aggregated data analysis approach, assessing latent differences in a scale-level dependent variable using a nominal-level variable described by two or more categories. As for correlation and MLR, these approaches also rely on a normal distribution assumption of variables within groups, all described by equal standard deviations. To overcome this limitation, other non-parametric alternatives might be implemented such as the Kruskal–Wallis H test proposed in [14]. Nonetheless, every aggregation technique describes whether a contextual partition is significantly correlated to a specific variable losing fine-grain description at the individual level, ultimately leading to interpretational ecological fallacy biases [42]. On the contrary, in disaggregation techniques [43] each feature under analysis is labelled using n dummy variables that describe the association to the region in which the individual is located; however, the linear relationship between the dependent variable and the regressors is not affected. This technique only allows the intercept value to be adjusted, and assumes the same relationship exists between variables in every group.

To overcome these limitations, regressions should be separately replicated for each sub-region under analysis, allowing the different variable relationships occurring in different regions to be explored. Some of the aforementioned works adopted regression via subgroup solutions [16]; however, separation approaches lack an assessment of inter-class variability.

A fourth approach that should be mentioned here is multilevel linear modelling (MLM). This approach, traditionally implemented in social studies, allows variables defined at different aggregation levels (often administrative units) with a nested structure to be investigated [38]. When considering our research design, contextual partitions identified areas of similar morphological properties, but no descriptors were associated with these higher-level aggregations.

Bearing in mind the aforementioned observations, individual regressions were implemented that allowed specific solutions for both model and variable selection procedures to be explored for each sub-region.

2.4. Multicollinearity and Variable Selection Procedure

The proliferation of studies investigating different urban form features and methodological approaches in relation to micro-retail distribution has resulted in a rich yet fragmented literature. Despite evidence about the individual importance of specific aspects of urban form on micro-retail distribution, an overall picture of the role of the urban built environment is still missing. This same observation might be found in the origin of a recent trend in urban form literature (beyond the specific case of the micro-retail) interested in bridging and analyzing the combined effect of several urban form aspects and measures [44]. However, assessing the combined and relative importance of a large number of strongly correlated urban form descriptors comes into conflict with the assumption of independent variables underlying traditional regression approaches. This limitation is even more evident in the current work, where several variables are specifically conceived for the detection of aspects of urban form that are different but still correlated.

Although multicollinearity does not influence overall model precision, its main consequences concern the analysis and interpretation of individual regression coefficients, preventing isolation of the individual contribution of each explanatory variable [45,46]. In order to detect and reduce multicollinearity issues, several approaches have been proposed and traditionally applied.

Bivariate correlation coefficients and tolerance-based diagnosis (i.e., variance inflation factor-VIF) represent two traditional approaches that allow the regressors at the origin of multicollinearity issues to be identified. Bivariate correlation coefficients require an expert-based selection, with subsequent concerns about the robustness/reproducibility of the procedure; moreover, the evaluation of every couple of variables becomes a highly time-consuming procedure with large datasets. On the contrary, stepwise routines have been elaborated for tolerance-based diagnosis. Although these procedures support and automatize the process of variable selection [47], they also have been demonstrated to be sensitive to small perturbations in initial data [48], and to produce biased regression coefficients [49]. Both correlation and tolerance-based approaches do not consider dependent variables to be a targets of the selection process, and only explore the intercorrelation between regressors [50]. To overcome this limitation, Sevtsuk [23] implemented a variable selection based on a statistical significance threshold that was applied to each predictor regression coefficient.

A second approach to dimensionality reduction includes procedures such as factor analysis, linear discriminant analysis and principal component analysis. These approaches identify lower numbers of unobserved variables called factors, which are expressed as linear combinations of higher numbers of correlated variables. In the specific case of micro-retail distribution, factor analysis is implemented (e.g., [26]). Despite the mathematical similarities between several available methods, different results might arise with possible complications in the interpretation of each factor [51]. The main issue here is that the direct interpretation of original features is lost; moreover, different variable aggregations in each sub-region hinder any intra- and inter-level comparative analysis within our partition.

In the present work, we prioritized the identification of variables that are objectively measured and individually observable in their disaggregated forms, rather than subjective/latent/composite factors, in order to facilitate both the interpretation of each individual indicator (or group of indicators) and comparative analyses of the different sub-regions of the study area.

To meet these goals, we implemented penalized regression (PR) [49] procedures in the present work. PR is a recent feature-selection approach that allows identification of the most significant subsets of features of a targeted variable by removing features characterized by low relevance and high redundancy [52]. Using a computationally efficient procedure, PR reduces the original model complexity to a simpler, final model that encompasses the most significant variables.

Although PR has recently been applied in micro-retail-related studies [53–55], to the best of our knowledge it has not been implemented and assessed within urban morphology. Implementing PR in our case study allowed us to deal with multicollinearity in our dataset, and to achieve our third goal of outlining the subset of individual urban morphological variables most related to micro-retail spatial distribution within each sub-region under analysis.

2.5. Objective

In the previous sections, we discussed the main analytical approaches for analyzing the relationship between micro-retail distribution and urban form. To summarize, a simple bivariate correlation analysis is the most adopted approach used in urban studies when investigating the individual relationships of single variables with micro-retail distributions. However, when the focus of analysis combines several urban environment indicators, MLR has been proposed as a superior alternative.

The intrinsic statistical characteristics of our variables might represent an important restriction affecting the underlying assumptions of both bivariate correlation and MLR and, consequently, the validity of their outcomes. We highlighted their conceptual and methodological limitations when modelling the discrete, non-negative, highly skewed nature of micro-retail distribution. While the absence of stores should be considered an integral part of the process describing store distribution, the resulting zero-inflation is traditionally overlooked or manually removed. Increases in the number of urban variable descriptors and the spatial extents needed to deal with multicollinearity among a large set of independent regressors require different values depending on the morphological context under study. These two aspects are still overlooked or discussed individually without a common methodological framework; nonetheless, only their combined evaluation can reveal important information on the roles played by each urban form aspect on micro-retail distribution.

Based on these observations, the following sections will show how the combination of GLM and PR approaches represents a better alternative to MLR models. This well-established modelling approach is able to deal with the four aforementioned aspects within a coherent, robust and innovative methodological framework.

3. Materials and Methods

In this section, we present the study area and databases underlying both urban and micro-retail descriptors. Next, the spatial unit of analysis and different families of street-based urban form descriptors are briefly defined. Finally, the model and variable selection procedures are described.

3.1. Case Study and Data Sources

The analytical protocol proposed in this work was tested on a real case study of the French Riviera metropolitan area in southern France. This polycentric coastal settlement comprises 88 municipalities that are structured around six main urban centers. From west to east we find: the Cannes–Grasse–Antibes conurbation, with 74,200, 51,000 and 73,800 inhabitants in their central cities, respectively; Nice, with 343,000 inhabitants, representing the largest municipality of the French Riviera and its administrative center; and the enclave of Monaco and the border city of Menton, with 38,000 and 28,000 inhabitants, respectively. Within these six municipalities about 70% of all micro-retail

businesses is found. Spread around these main centers, 295,000 people live in smaller cities, villages and hamlets surrounded by vast residential areas, according to the morphological properties of the site. All these differently sized centers are interconnected by a pervasive, discontinuous and car-dependent residential fabric. With a total of more than 1 million inhabitants, the French Riviera is the seventh most populated metropolitan area in France.

The combination of all these elements produces a sequence of urban centers and peripheral areas of different sizes that encompass a large variety of urban forms. Previous studies have disentangled the high heterogeneity of the study region, identifying typo-morphological regions both at district and neighborhood scales [56,57]. These sub-regions correspond to different urban morphological contexts characterized by specific combinations and distributions of urban configurational and morphological descriptors; moreover, for each of these regions, different zero-inflation and overdispersion properties of the micro-retail distribution are also observed. These characteristics allow the present work to overcome the limitations of traditional works that have investigated only individual core regions of medium- or large-sized monocentric cities [10], and to assess the current analytical procedure under different contextual and statistical conditions.

Two sources of data are considered in this work. The official data about micro-retail distribution is provided by the local Chamber of Commerce of Nice Cote-d'Azur (CCINCA), counting about 50,000 businesses and services active as of 1 January 2017. (More recently, this same information has been made available at the national level by the national statistics agency (INSEE).)The address information allowed us to geocode the database and provide a spatial representation of the phenomena under study. This process was realized through the National Open Addresses Database (Base d'Adresses Nationale Ouverte (BANO)). The BANO geolocation tool associated a score of the geocoding results describing the localization precision at four levels: null, municipality, street and house number. From our original dataset: (i) 7% of the data presented missing information, or fell outside of our study area, and was thusly excluded from our analysis; (ii) 2% of information was geo-localized at the municipality level and 13% was at the street level—the cause of these mis-localisations was often a result of incomplete address information in the original database such as missing civic number, misspelt street name, incorrect name of an isolated hamlets and so on, and a manual correction was carried out when the correct retail activity address was available from other online sources; (iii) 78% of data were correctly located at the house-number level. We obtained a final dataset of 45,726 stores distributed across 33,221 locations (several activities shared the same addresses), 82% with a precise civic number and 18% at the street level (positioned at street segment midpoints). In 135 locations, large planned centers were found with retail surfaces higher than 2000 square meters. This specific retail format does not possess the same combinations of locational factors as smaller activities [58], however its presence has the potential to profoundly modify the surrounding urban morphology and flow, making these centers an attractive element for smaller activities (i.e., retail locomotives). For this reason, these activities (from now on named "anchor stores"), were excluded from the original dataset and considered as a locational factor for smaller commercial activities (see Section 3.2).

Urban form descriptors were based on the geographic databases (BD TOPO, 2017) from the French National Institute of Geographical and Forest Information (IGN). Four layers of urban morphological elements were used: building, street-network, parcel and digital terrain model (DTM).

Based on these data sources, well-established GIS-based protocols were implemented for the elaboration of the different urban morphological descriptors, while statistical procedures were implemented with R libraries [59]. The use of relatively simple data and available analytical/statistical protocols make this work reproducible for future comparative studies.

3.2. The Variables under Investigation

The spatial unit of analysis was the street segment. Streets represent one of the most used spatial units, and have been attracting attention in the last 20 years from urban designers, configurational

studies, morphologists and urban geographers [60]. Streets are considered to be the bridging element between different methodological and theoretical approaches [44].

The street segment is here defined as the centerline between two street-junctions. Four reasons motivate this choice, the first of which being that "the dominant network model is the one that represents the street junctions as vertices in the graph and the linear street segments as its edges" [61]. Secondly, by using street network centerlines, the primary approach allows the independent identification of configurational properties according to the physical shapes and sizes of built forms surrounding street segments (isolating configurational properties of the network from morphometric measures of the streetscape and fabrics). Thirdly, the use of a centerline permits a geometrical reference when studying streetscapes from the street point of view (measures of setback, parallelism of facades and so on are used as reference street edges and/or street centerlines). The street segment therefore becomes both a geometrical (streetscape measures, the geometry of retail agglomerations, etc.) and metric (local configurational properties, local morphological patterns, etc.) reference [62], and the use of visual axes as in SSx or alternative street-like representations of the street network provide a distorted reference system for streetscape descriptors. Finally, the street segment represents a behaviorally oriented partition of space, which is better suited for socioeconomic phenomena such as the distribution of retail businesses in urban space [35].

To describe different aspects of urban form, several computer-aided procedures from established scientific literature were implemented for our study region. Each street segment was characterized by more than 100 street-based descriptors of urban form (further details about urban form indicators are described in Appendix B).

Four main subsets of indicators can be recognized: the first comprises 40 indicators that have been defined to describe street network configurational properties using the MCA protocol [8,9]. Local Reach, Straightness, Closeness and Betweenness centralities are assessed at different scales and impedances on pedestrian and vehicular modelled street-networks (300-, 600- and 1200-meter radii and 5- and 20-minute radii, respectively. Their normalized versions are obtained following a two-step floating catchment area procedure (2SFCA) [63].

The second subset of indicators is made up of 36 indicators describing the street-network accessibility towards public squares, coastline and anchor stores, which are considered influential components of an urban form on micro-retail distribution. As with the previous metrics, several scales and impedances were considered.

From the urban design and urban morphological literature, 30 indicators describing the built form layout along the street edges have been defined (also named skeletal streetscapes [64]). Several GIS protocols have been proposed in recent urban form literature [64–67]. and indicators such as façade alignment, building set-back, average building height and so on are calculated while considering building distribution within a 50-meter distance from street edges through the definition of street-based proximity bands (PBs) and sightlines [56,67].

Finally, street-based contextual variables/partition have been obtained through the implementation of the Multiple Fabric Assessment procedure [56], wherein each street segment is associated with nine values, with each one describing the probability of association with different urban fabric types. In more central and compact regions, historical centers, traditional planned fabrics with adjoining buildings and discontinuous fabrics of buildings and houses are found (respectively, UF1–3). Semi-peripheral and peripheral regions are prevalently composed of modernist urban fabrics and suburban areas with lower/higher natural constraints (respectively, UF4–6). Finally, the least dense regions are described by connective artificial fabrics and natural spaces of hills and mountains (respectively, UF7–9). This urban fabric partition is illustrated in Figure 1 and further described in [57]. The study of the spatial organization of these nine urban fabrics allows the identification of three morphological macro-regions within a metropolitan area: First-, Second-Age City (following the morphological categories of [68]) and Natural Space. These two typo-morphological partitions of the study area, illustrated in Figure 1, define the sub-regions where count regression approaches are individually applied; the limited number

of streets with stores within the Natural Space and UF7–9 prevent the implementation of our analytical procedures in these specific morphological regions.

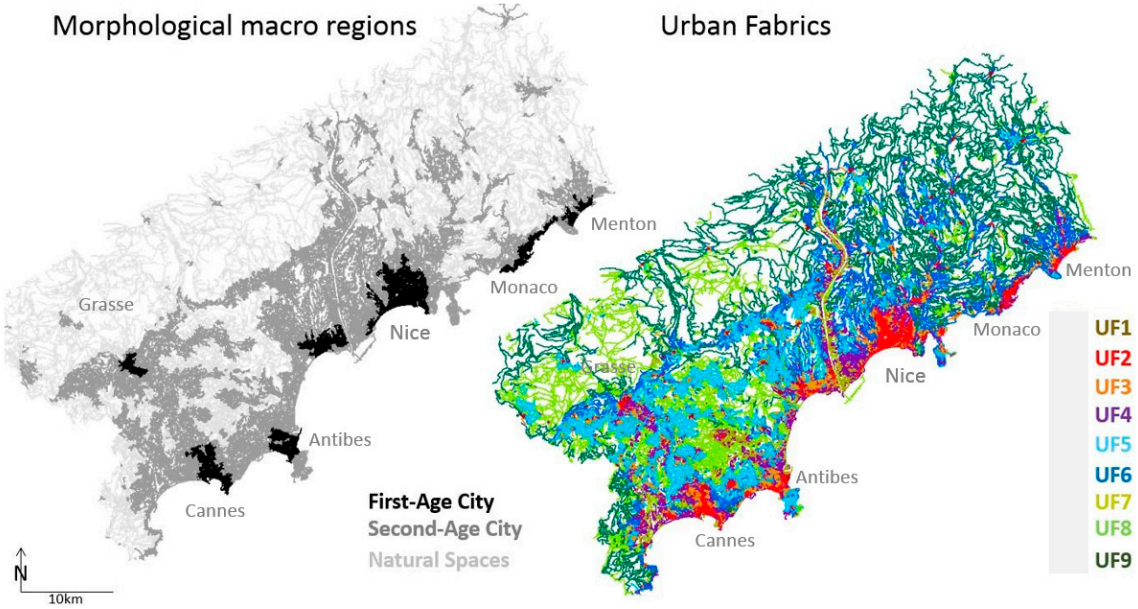

Figure 1. The French Riviera study region: the three morphological regions (First-/Second-Age City and Natural Spaces (left)) and the nine urban fabrics (UF1–9 (right)). Source: [56,57].

Of the almost 100,000 street segments composing the whole street network of the French Riviera, we focused on those where built-up elements were found within 50 meters from street edges. Streets crossing natural areas, large public parks and small connective segments were excluded, reducing our dataset to 63,071 units. Each street segment was defined by the number of small stores representing the target variable of our models. Different values of zero-inflation, street density and overdispersion were observed in each morphological sub-region (Table 1).

Table 1. Micro-retail distribution, variance and zero-inflation values for the overall study region and within each morphological context.

	Global	1st-A.C.	2nd-A.C.	Natural	UF1	UF2	UF3	UF4	UF5	UF6	UF7	UF8	UF9
N° streets with build.	63071	14143	43969	4959	5523	8030	7210	10310	16528	9032	2018	2543	1877
% streets UF(i) with build.	63.1	85.3	72.5	23.3	93.1	87.6	90.2	68.3	84.4	74.1	27.5	24.2	17.6
[%] Streets with Retail	22.6	36.9	14.3	5.5	25	48	18	19	11	10	14	7	5
Avg. Retail Street Count	0.66	1.81	0.35	0.08	0.69	2.8	0.37	0.63	0.17	0.14	0.64	0.13	0.06

Before proceeding with a description of the modelling protocol, two further aspects should be underlined. Firstly, the same four limitations presented in Section 2 still persist when using other fine-grained spatial unit definitions and urban form descriptors. As such, the modelling solution presented in this paper might also be tested and implemented with other street-based spatial unit definitions (i.e., axial streets, named streets, raster-based solutions, plots, etc.). Nonetheless, the combination of several urban form analytical procedures, each one based on ad hoc spatial unit definitions, would require a supplementary artificial manipulation of the variables, which would lead to the introduction of a statistical bias and compromise both the modelling and variable selection procedure performances and outcomes.

Secondly, this work focuses on the study of the physical properties of urban form, and does not take into consideration any socioeconomic and land-use regulation aspects. It is fully recognized that such aspects play an important role as locational factors in retail distribution, and are each related to urban form in different ways. For this reason, both modelling performance measures and variable

selection procedure could be strongly dependent on these variables, confounding the role of other urban descriptors. Their exclusion from the modelling procedure allows the roles of different properties of the urban built environment to be explored and pointed out. Further research would be needed to disentangle the roles of urban form, socioeconomic aspects and planning constraints.

3.3. Modelling Micro-retail Distribution: From Linear to Count Regression Approaches

As discussed in the previous section, count regression approaches seem to be best suited to our case study. These methods have been widely developed over the last 50 years [30,69–72]. GLMs have been specifically developed to handle count data: a mathematical transformation on the dependent variable is operated, considering the true distribution of errors and assuming a distribution from an exponential family (i.e., binomial, Poisson, multinomial, etc.). A linear relationship is then investigated between the independent variables and the transformed response rather than its raw values. A maximum likelihood estimation (MLE) procedure is implemented for the estimation of the model parameters.

When the distribution of the dependent variables (and errors) follows a Gaussian (G) distribution, the identity function describes the transformation and, subsequently, the GLM results in the same estimates as the traditional MLR [72]. When the variable to be analyzed is represented by a count variable, the random component assumes the form of a Poisson distribution and the corresponding transformation is usually a log function. The resulting model is called a log-linear or Poisson regression model (P). However, the main assumption of a Poisson model is that the mean and standard deviations of the observed dependent variable are equivalent, an assumption that is not met when the dependent variable is characterized by high heterogeneity. Negative binomials (NBs) might be considered an alternative to the Poisson model, and this specific form provides a built-in solution to account for overdispersion. P and NB represent two interesting alternatives to G/MLR overcoming the restrictive assumption of homoscedasticity while considering the true distribution of errors.

Despite being able to handle discrete non-negative and skewed distributions, the models presented so far cannot handle overdispersion due to zero-inflation (heuristic rules suggest a presence of zeroes not higher than 20% of the expected values, which is far less than what was observed in our target variable). In such situations, the GLM approach proposes alternative solutions that are able to integrate and model an excessive presence of zeroes.

With zero-inflated (ZI) regression models [37], zeros originate according to two simultaneous processes. The probability distribution of zero-inflated models are defined as the combination of a logistic part modelling the structural zeros (or true zeros) and a count part assuming a P (ZIP) or NB (ZINB) form from which random zeros (or false zeros) are produced.

Zero-alternated (ZA, or hurdle) approaches [73,74] model all zeros as one part, while the non-zero part is modelled with zero-truncated count regressions. The implementation of the P or NB forms into the zero-truncated part of the model result in zero-alternated Poisson (ZAP) and negative binomial (ZANB) models.

Implementing ZI and ZA models allowed us to explore the possibility that two processes might determine the observed zero and non-zero values instead of considering that these values come from the same data-generating process. Both ZI and ZA are described by the combination of logistic regression and Poisson (ZIP-ZAP) or negative binomial (ZINB-ZANB) models. The main difference among these approaches is that the former considers the observed distribution of values to be the result of the combined processes with a possibility of distinguishing between structural and random zeros, while the latter supposes two separate generating processes producing zero and non-zero values. Finally, the opportunity to use P and NB both in ZI and ZA allows us to control for the combined overdispersion of count and zero parts.

For the three models previously described (G, P, NB), four additional models were implemented and compared (ZIP, ZAP, ZINB, ZANB). The seven models here presented were performed on the overall study area and eight aforementioned sub-regions.

GLM is a powerful technique that enables a wide number of modelling approaches beyond the traditional MLR to investigate different aspects of the dependent variable statistical distribution. While the implementation and comparison of these approaches have been already discussed in several disciplines, no work has investigated this specific aspect in the case of micro-retail distribution and urban form. The implementation of a comparative analysis of seven regression models allowed us to understand whether specific processes should be considered when describing the relationship between urban form and micro-retail distribution. Goodness-of-fit measures are described in the next section as support for the model selection procedure.

Before proceeding with further specifications, another observation should be made. Micro-retail distribution is frequently measured as a density; one might argue that the raw count of stores might be strongly biased by the size of the underlying spatial unit. A specific approach to handle density variables is possible when implementing GLM. Density might be seen as a rate between a count value (the store number) and the underlying spatial unit size (street length), also named the exposure variable. GLM handles exposure variables using simple algebra, changing the dependent variable from a rate into a count by simply multiplying both sides of the equation according to the exposure variable and moving it to the right side of the equation. In the final model, the exposure variable becomes a term of the regression coefficients, also called the offset variable. With this solution, GLM permits the preservation of the natural form of the counting data, which accounts for the variabilities determined by the underlying spatial unit dimension.

3.4. Modelling Selection: Goodness-of-fit Measures

Defining a common procedure by which to assess and compare the different models is a task of paramount importance when identifying the most adapted modelling approach.

Since the traditional coefficient of determination R2 requires a homoscedastic distribution of error, extensive scientific literature has focused on pseudo-R2 for count regression models [75–78]. Nonetheless, there is no consensus on which measure should be preferred, and each choice might lead to certain drawbacks [79]. For example, goodness-of-fit measures have been specifically conceived for each type of GLM regression, preventing their application in a large variety of models with the final goal of supporting the model selection phase.

To overcome this limitation, measures based on information criteria (IC) have become increasingly popular. The notion behind IC approaches is the need to find a compromise between likelihood maximization and the principle of parsimony, which favors simpler models [72]. The Akaike information criterion (AIC) [80] is obtained as $AIC = 2K - 2\log(L(\hat{\theta}|y, M))$, where K is the number of estimable parameters that correspond to the degree of freedom, and $L(\hat{\theta}|y, M)$ is the maximum value of the likelihood function for the model M. In other words, the AIC score is an estimate of a constant based on the degrees of freedom of a model, plus the negative log-likelihood of the model knowing the data. A lower AIC score reflects models that are closer to reality. AIC scores do not have a specific meaning when independently considered, but a comparison of AIC scores from different models can help an analyst rank and select the best solutions from a finite set of models. An AIC can only be obtained from GLM approaches that allow non-nested models to be compared, which ordinary statistical tests cannot do.

The implementation of likelihood ratio-based tests (LR-test) provides an analyst with further evidence highlighting statistically significant differences between IC scores. The null hypothesis of an LR-test is whether both compared models are equally close to the true model. If the null hypothesis is not verified, one of the two models should be considered as having a better performance. The Vuong test [81] for non-nested models is so far the most applied LR-test among the different domains of the scientific literature without any restrictions on GLMs. In this work, AIC scores and the non-nested Vuong testing were used to quantify and rank our model performances and, ultimately, guide the model selection. As we were aware of possible biases when considering ZI models [82], rootgrams [83] were also implemented as a graphic solution to support the model assessment.

While the aforementioned procedure assessed and supported the model selection procedure, two additional aspects should be outlined. Firstly, loglikelihood-based measures allowed comparison only if models shared the same underlying dataset (both in terms of variables and records). Therefore, the same approach was not suitable when comparing global model outcomes with those obtained from the subgroup regressions approach. Secondly, AIC is a global measure, and does not allow to appreciate the roles of overdispersion and zero-inflation on model performance outcomes.

Other parameters were also implemented, allowing the description of different aspects of the model outcomes. Count pseudo-R2 [84] was implemented as the proportion of correct estimates on the overall number of predictions; similarly, weighted accuracy, recall and F1 scores were also provided. Traditional measures of dispersion of the residuals (mean absolute and standard deviation) for each model completed the model outcome description. These measures were applied while considering zero and count parts of each model separately, thus revealing their relative impacts on the overall goodness-of-fit measures.

3.5. Feature Selection.

In the previous sections we defined a model selection procedure to identify the most adapted approach to describing micro-retail distribution, which we based on overall goodness-of-fit measures, without considering the specific combination of regressors. Nonetheless, as outlined in Section 2.4, non-experimental studies are nearly always characterized by the presence of multicollinearity; this was even more true in this work, where different facets and metrics of the same phenomenon—the urban physical form—were studied and combined. Another goal of this work was to outline the subsets of individual urban morphological variables related to micro-retail spatial distribution within each sub-region under analysis.

In order to achieve this objective, a specific category of feature selection—penalized regression (PR)—provided a built-in solution for GLM count regression approaches. While the goal of traditional selection procedures is to remove predictors from a model that are not considered significant and thus set their regressor coefficients to zero, the idea underlying PR is to penalize them toward zero without forcing them to be exactly zero (for this reason, these methods are also known as shrinkage or regularization methods). In this way, the complexity of the model is reduced while keeping all or part of the variables in the model. PR traditionally requires the choice of a shrinkage value of lambda to define the magnitude of the penalization.

Three main penalized regression procedures are most commonly used: ridge, least absolute shrinkage selection operator (LASSO) and elastic net (Enet). In ridge PR, the loss function underlying the regression models is augmented to minimize the sum of the squared residuals while taking into account and penalizing the size of the parameter estimates, with the final goal of shrinking them toward zero. In LASSO PR [49], the regression coefficient to be shrunk toward zero as well as those with a minor contribution might be forced to be exactly equal to zero. Two different penalization functions are considered in ridge and LASSO approaches. While ridge seems to be more frequently adapted when coefficient parameters are of a similar size, LASSO regression is typically adapted when a model presents a subset of variables with high coefficient parameters while the remaining have very small coefficients [85].

Finally, Enet regression combines both Ridge and LASSO penalization approaches, allowing both the coefficient to shrink toward zero while also setting some variables to equal zero precisely, producing simpler and more interpretable models. Implementing Enet regression in our case study enabled us to outline the subset of urban morphological variables most related to the spatial distribution of retail.

In order to find the optimal values for the shrinkage parameters, specific iterative processes were implemented from a large number of possibilities using optimization procedures based on IC such as AIC or, similarly, the Bayesian Information Criterion (BIC, [86]). For each study region, we asked the Enet algorithm to explore 20 values of lambda. The regression coefficients reported in this work correspond to the penalized model for which the lowest BIC scores were observed.

4. Results: Application to the French Riviera Case Study

The outcomes of the procedure previously described are herein presented as follows. First, we focus on the model selection outcomes. Since the overall model selection criteria and predictions are not influenced by multicollinearity problems, the role of individual regressors is temporarily overlooked. Once the most adapted modelling procedure is defined, the second part of this section is dedicated to the results of the variable selection procedure.

4.1. Model Selection

Seven regression models (G, P, NB, ZIP, ZINB, ZAP, ZANB) were implemented on the overall space, on two sub-regions at the district scale (First/Second-Age City) and on six urban fabrics (UF1–6). Each of the 63 models is described in Table 2 according to the following set of four descriptors: AIC, -2loglikelihood, number of features (streets) and number of parameters c (variable number + number of parameters of the model). The best model was found to correspond with the lowest AIC value.

Table 2. Model selection for the overall study area and each morphological region.

		Model Selection						
		G	P	NB	ZIP *	ZINB **	ZAP *	ZANB *
Global	AIC	284,770.9	120,492	90,024.06	100,454.9	87,373.54	101,170.5	88,247.1
	−2log-likelihood	284,558	120,282	89,812	100,034	86,956	100,750	87,824
	n	63,071	63,071	63,071	63,071	63,071	63,071	63,071
	c	105 + 2	105 + 1	105 + 2	210 + 2	210 + 3	210 + 2	210 + 3
					**	**	*	*
First	AIC	77,041	47,955.88	36,477.71	41,594	35,451	41,729	35,869
	−2log-likelihood	76,828	47,696	36,264	41,150	35,036	41,280	35,450
	n	14,143	14,143	14,143	14,143	14,143	14,143	14,143
	c	101 + 2	101 + 1	101 + 2	202 + 2	202 + 3	202 + 2	202 + 3
					*	**	*	*
Second	AIC	171,475	68,000	49,808	54,969	48,841	55,431	49,198
	−2log-likelihood	171,260	67,862	49,616	54,522	48,416	54,976	48,746
	n	43,969	43,969	43,969	43,969	43,969	43,969	43,969
	c	102 + 2	102 + 1	102 + 2	204 + 2	204 + 3	204 + 2	204 + 2
					**	*	*	*
UF1	AIC	22,254	10,816	9405	9883	9137	9968	9480
	−2log-likelihood	22,078	10,774	9280	9580	8968	9662	9164
	n	5506	5506	5506	5506	5506	5506	5506
	c	93 + 2	93 + 1	93 + 2	186 + 2	186 + 3	186 + 2	186 + 3
					**	**	*	*
UF2	AIC	47,155	37,145	27,049	31,622	26,158	31,721	26,352
	−2log-likelihood	46,948	36,914	26,866	31,218	25,786	31,308	25,976
	n	7954	7954	7954	7954	7954	7954	7954
	c	94 + 2	94 +	94 + 2	188 + 2	188 + 3	188 + 2	188 + 3
					*	*	*	*
UF3	AIC	22,097	9976	9025	9161	8909	9267	9058
	−log-likelihood	21,890	9766	8830	8760	8522	8890	8662
	n	7108	7108	7108	7108	7108	7108	7108
	c	93 + 2	93 + 1	93 + 2	186 + 2	186 + 3	186 + 2	186 + 3
					*	**	*	*
UF4	AIC	48,306	25,602	16,408	19,036	16,023	19,154	16,135
	−2log-likelihood	48,108	25,340	16,186	18,598	15,622	9357	15,730
	n	10,259	10,259	10,259	10,259	10,259	10,259	10,259
	c	94 + 2	94 +	94 + 2	188 + 2	188 + 3	188 + 2	188 + 3
					*	*	*	*
UF5	AIC	36,232	15,636	13,765	14,156	13,691	14,267	13,778
	−2log-likelihood	36,138	15,462	13,596	13,800	13,316	13,946	13,458
	n	16,453	16,453	16,453	16,453	16,453	16,453	16,453
	c	86 + 2	86 + 1	86 + 2	86 + 2	86 + 3	86 + 2	86 + 3

Table 2. *Cont.*

	Model Selection							
					*	*	*	*
UF6	AIC	16,590	7186	6721	6793	6638	6883	6800
	−2log-likelihood	17,878	7358	6720	6714	6586	6792	6658
	n	8889	8889	8889	8889	8889	8889	8889
	c	86 + 2	86 + 1	86 + 2	86 + 2	86 + 3	86 + 2	86 + 3

Note: Red and green colors highlight higher and lower AIC scores, respectively. AIC, Akaike information criteria; c, number of parameters in the model; n, feature number; G, Gaussian; P, Poisson; NB, negative binomial; ZI, zero-inflated; * and ** indicate that the AIC was obtained with logit and probit functions, respectively.

Higher values of AIC were found, as expected, corresponding to the linear (G) and Poisson (P, ZIP, ZAP) regression models. The inappropriateness of these models is here empirically confirmed independent of the spatial region under analysis. Lower AIC scores were found for the NB, ZINB and ZANB models. Despite small differences among these three approaches, ZINB always presented the lowest AIC values.

As can be observed in Table 3 where non-nested Vuong test results are reported, the statistically significant superiority of ZINB was confirmed in every region with the exception of UF5 and UF6. In these two specific cases, $p > 0.05$ when comparing the AIC values of ZINB and ZANB. These outcomes provide solid evidence regarding the presence of a double process defining micro-retail distribution.

Table 3. Results of the Vuong LR-test between our seven models for the overall study area and each morphological region.

Global	Tested Models	G vs. P	P vs. NB	NB vs.ZIP	ZIP vs. ZINB	ZINB vs. ZAP	ZINB vs. ZANB	ZINB vs. ZINB
	Vuong test statistic	−89.546	−21.39	13.692	−16.887	17.724	9.313	w2 = 0
	p	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	0.5
	Best Model	P	NB	ZIP	ZINB	ZINB	ZINB	-
First	Tested Models	G vs. P	P vs. NB	NB vs.ZIP	NB vs. ZINB	ZINB vs. ZAP	ZINB vs. ZANB	ZINB vs. ZINB
	Vuong test statistic	−37.51	−19.416	11.817	−18.483	15.491	7.93	0
	p	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.2 \times 10^{-15}$	$<0.713 \times 10^{12}$	0.5
	Best Model	P	NB	NB	ZINB	ZINB	ZINB	-
Second	Tested Models	G vs. P	P vs. NB	NB vs.ZIP	NB vs. ZINB	ZINB vs. ZAP	ZINB vs. ZANB	ZINB vs. ZINB
	Vuong test statistic	−66.592	−14.391	8.313	−14.731	11.29	4.46	−8.854
	p	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.74 \times 10^{-8}$	$< 0.22 \times 10^{-15}$	$<1.296 \times 10^{-6}$	$< 0.22 \times 10^{-15}$
	Best Model	P	NB	NB	ZINB	ZINB	ZINB	-
UF1	Tested Models	G vs. P	P vs. NB	NB vs.ZIP	NB vs. ZINB	ZINB vs. ZAP	ZINB vs. ZANB	ZINB vs. ZINB
	Vuong test statistic	−40.545	−9.674	3.179	−9.66	7.186	4.663	0
	p	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	0.7×10^{-3}	$<0.2 \times 10^{-15}$	3.34×10^{-14}	1.56×10^{-6}	0.5
	Best Model	P	NB	NB	ZINB	ZINB	ZINB	/
UF2	Tested Models	>G vs. P	>P vs. NB	>NB vs.ZIP	>NB vs. ZINB	>ZINB vs. ZAP	ZINB vs. ZANB	>ZINB vs. ZINB
	Vuong test statistic	−13.339	−17.263	10.96	16.99	14.338	5.633	0
	p	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	8.85×10^{-9}	0.5
	Best Model	P	NB	NB	ZINB	ZINB	ZINB	/
UF3	Tested Models	G vs. P	P vs. NB	NB vs.ZIP	ZIP vs. ZINB	ZINB vs. ZAP	ZINB vs. ZANB	ZINB vs. ZINB
	Vuong test statistic	−49.992	−9.183	−1.04	-5.59	6.9	4.357	0.385
	p	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	0.0856	$<0.22 \times 10^{-15}$	4.42×10^{-12}	2.18×10^{-5}	0.5
	Best Model	P	NB	ZIP/NB	ZINB	ZINB	ZINB	/
UF4	Tested Models	G vs. P	P vs. NB	NB vs.ZIP	NB vs. ZINB	ZINB vs. ZAP	ZINB vs. ZANB	ZINB vs. ZINB
	Vuong test statistic	−22092	−10.486	6.835/7.436	−8.246	8.558	2.103	0.001
	p	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	0.4×10^{-12}	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	0.0178	0.987
	Best Model	P	NB	NB	ZINB	ZINB	ZINB	/
UF5	Tested Models	G vs. P	P vs. NB	NB vs.ZIP	NB vs. ZINB	ZINB vs. ZAP	ZINB vs. ZANB	ZINB vs. ZANB
	Vuong test statistic	−52	−8.314	2.19	−9.16	6.505	1.97	0
	p	$< 0.22 \times 10^{-15}$	$< 0.22 \times 10^{-15}$	0.0153	$<0.22 \times 10^{-15}$	3.87×10^{-11}	0.9756	0.5
	Best Model	P	NB	NB	ZINB	ZINB	ZINB/ZANB	/
UF6	Tested Models	G vs. P	P vs. NB	NB vs.ZIP	NB vs. ZINB	ZINB vs. ZAP	ZINB vs. ZANB	ZINB vs. ZANB
	Vuong test statistic	−43.096	−6.158	−0.187	−5.8/15.64	1.95/15.64	1.85	1.2
	p	$< 0.22 \times 10^{-15}$	3.69×10^{-10}	0.425	2.06×10^{-9}	2.3×10^{-2}	0.107	0.107
	Best Model	P	NB	NB	ZINB	ZINB	ZINB/ZANB	/

Note: G, Gaussian; P, Poisson; NB, negative binomial; ZI, zero-inflated.

These observations are further confirmed when plotting the relative rootgrams. In Figure 2 we might observe rootgrams for the seven models implemented on the overall study area, and similar behaviors can be observed for every morphological sub-region. Linear (G) and Poisson regression

models (P, ZIP, ZAP) did not account for overdispersion, contrary to all negative binomial regression models (NB, ZINB, ZANB).

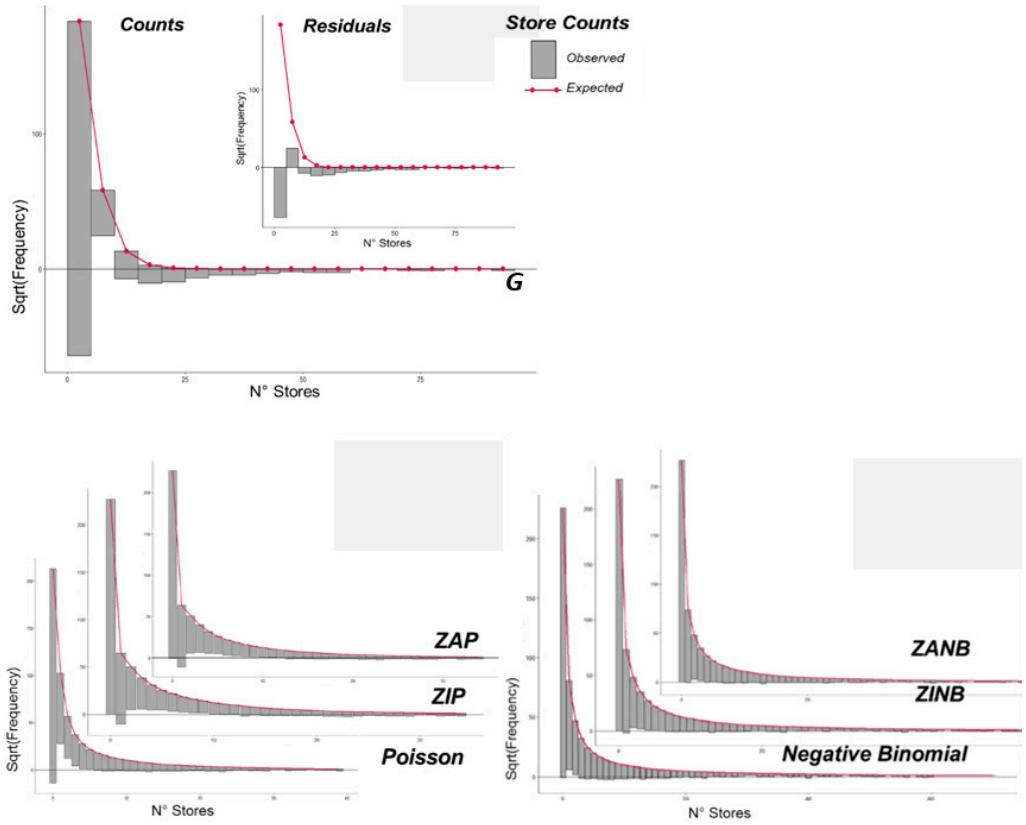

Figure 2. Rootgram of the global model (overall space study and variables) for: linear model G (top), Poisson regressions P-ZIP-ZAP (bottom-left) and negative binomial regressions NB-ZINB-ZANB (bottom-right). The last group of models showed a better fit between expected and observed values.

Having determined ZINB to be the modelling approach that best fit our study case, we can now observe the impact of regression according to the different subgroups: Table 4 gathers the set of 13 measures previously described using the count and zero outcomes for each sub-region under analysis. The lowest accuracy and sensitivity values were found in the more central compact fabrics. Inversely, peripheral urban fabrics showed higher values. Precision relatively to the count parts dropped in peripheral urban fabrics where higher zero-inflation was observed.

When implementing ZINB models separately on the First- and Second-Age City partitions, the accuracy of the overall model improved +0.38% and +0.98%, respectively, while the accuracy level grew by +0.52% when using the six UFs. The decomposition of the overall study area showed minor improvements on the overall predictability of the model. However, different levels of improvement were observed when considering each sub-region individually: the accuracy of UF1–3 substantially improved by +4.66%, +12.29% and +2.55%, respectively, and the F1 score improved by +7.38%, 2.89% and 5.50%, respectively. As for UF5 and UF6, the accuracy was similar between the global and local models, while the F1 scores were higher in the latter. Only for UF4 did both accuracy and F1 scores show small variations between global and local models.

Table 4. Comparison of the results of ZINB models when global and sub-regions are evaluated for the same subgroup of features.

ZINB	*C*		*Sz*		*Sc*		*Pz*		*Pc*		*F1*		*E(T)*		*E(Tz)*		*E(Tc)*		*Sd(T)*		*Sd(Tz)*		Sd(Tc)	
	val	±[%]	val	±[%]	val	±[%]	val	±[%]	val	±[%]	val	±[%]	val	±[%]	val	±[%]	val	±[%]	val	±[%]	val	±[%]	val	±[%]
Global	0.716		0.839		0.178		0.907		0.134		0.536		0.624		0.228		2.350		1.997		0.707		3.943	
Global *	0.697		0.824		0.181		0.902		0.134		0.541		0.668		0.249		2.370		2.067		0.727		3.981	
Glob. (F + S)	0.704	0.98	0.834	1.15	0.177	−2.21	0.902	−0.03	0.136	1.13	0.547	1.06	0.662	−0.89	0.239	−4.02	2.380	0.39	2.091	1.15	0.751	3.16	4.018	0.93
First *	0.456		0.610		0.192		0.878		0.126		0.678		1.476		0.639		2.908		2.995		1.231		4.298	
First	0.484	5.78	0.654	6.78	0.192	-0.10	0.882	0.37	0.133	5.27	0.696	2.58	1.434	−2.93	0.574	−11.34	2.906	−0.09	2.998	0.08	1.248	1.32	4.273	−0.58
Second *	0.775		0.875		0.172		0.906		0.143		0.411		0.408		0.156		1.922		1.575		0.501		3.637	
Second	0.775	0.01	0.876	0.13	0.165	−3.86	0.905	−0.07	0.139	-2.84	0.409	−0.56	0.414	1.42	0.160	2.08	1.944	1.10	1.624	3.03	0.535	6.43	3.750	3.02
Global *	0.703		0.831		0.180		0.902		0.137		0.547		0.652		0.237		2.332		1.985		0.706		3.792	
MFA(6)	0.714	1.65	0.845	1.60	0.186	3.24	0.907	0.50	0.146	6.34	0.569	3.90	0.634	−2.80	0.224	−6.02	2.293	−1.68	1.967	−0.88	0.725	2.57	3.739	−1.41
UF1 *	0.628		0.756		0.246		0.888		0.171		0.584		0.685		0.301		1.833		1.512		0.615		2.496	
UF1	0.659	4.66	0.793	4.62	0.259	5.03	0.903	1.74	0.190	10.08	0.631	7.38	0.650	−5.42	0.261	−15.21	1.811	−1.20	1.649	8.35	0.621	0.95	2.812	11.22
UF2 *	0.288		0.400		0.167		0.876		0.106		0.728		2.209		1.102		3.395		3.747		1.563		4.874	
UF2	0.328	12.23	0.479	16.48	0.166	−0.94	0.885	1.02	0.111	4.84	0.749	2.89	2.113	−4.55	0.995	−10.80	3.312	−2.53	3.651	−2.65	1.652	5.38	4.681	−4.12
UF3 *	0.743		0.858		0.225		0.888		0.195		0.475		0.394		0.151		1.490		1.013		0.389		1.873	
UF3	0.763	2.56	0.881	2.63	0.228	1.36	0.892	0.40	0.216	9.91	0.503	5.50	0.380	−3.44	0.145	−3.78	1.443	−3.28	1.000	−1.27	0.510	23.76	1.722	−8.76
UF4 *	0.611		0.702		0.230		0.900		0.120		0.463		0.787		0.389		2.448		2.319		0.778		4.677	
UF4	0.615	0.76	0.707	0.62	0.236	2.56	0.902	0.23	0.124	3.20	0.469	1.24	0.777	−1.25	0.379	−2.65	2.439	−0.35	2.319	−0.02	0.756	−2.87	4.690	0.28
UF5 *	0.867		0.958		0.108		0.907		0.204		0.239		0.191		0.043		1.427		0.686		0.206		1.527	
UF5	0.868	0.13	0.959	0.09	0.112	3.05	0.910	0.27	0.204	0.05	0.266	10.34	0.192	0.41	0.045	4.70	1.417	−0.72	0.690	0.51	0.241	14.31	1.510	−1.15
UF6 *	0.886		0.971		0.100		0.916		0.225		0.248		0.156		0.029		1.326		0.546		0.169		1.119	
UF6	0.885	−0.03	0.967	−0.48	0.140	28.69	0.922	0.66	0.253	10.95	0.317	21.91	0.156	−0.14	0.035	16.85	1.270	−4.42	0.583	6.37	0.193	12.18	1.320	15.25

Note: For each goodness-of-fit measure the raw value (val) and percent change (±[%]) were measured between the model implemented on the overall space study (*) and for each sub-region. C, accuracy; Sc, sensitivity count part; Sz, sensitivity zero part; Pc, precision count part; Pz, precision zero part; F1, score; E(T), average tolerance; E(Tz), average tolerance zero part; E(Tc), average tolerance count part; Sd(T), standard deviation tolerance; Sd(Tz), standard deviation tolerance zero part; Sd(Tc), standard deviation tolerance count part.

When observing the separate sensitivity and precision values for the count and zero parts, an overall growth in precision could be observed, as well as a loss in sensitivity in the counting parts. Moreover, both accuracy and sensitivity improved for the zero parts, with the exception of UF5 and UF6. We might conclude that the decomposition of the study area in morphological subspaces (both UF and morphological macro-regions) improved the goodness-of-fit for traditional central areas, while semi-peripheral and peripheral regions seemed to be penalized. Since street elements in peripheral regions outnumbered those of more central areas, the model improvements achieved in central areas were diluted and reduced to modest values when evaluating the combined results of sub-models (First-/Second-Age City and UF1–6). These outcomes might support the hypothesis that urban form plays an important role in defining store distributions in compact traditional areas, while other locational factors should be considered for less dense, peripheral regions.

Projecting model outcomes in a geographical space allows the analyst to observe the spatial characteristics of the predictive power of the models. Ignoring these spatial representations of the model outcomes might hinder the detection of eventual model limitations, precluding important observations for future research developments. Specific patterns of residuals might suggest the omission of essential variables. In Figure 3, we illustrate the observed distribution of micro-retail, along with the predicted values and their errors, zoomed in on the city of Nice. From the left to the right we can compare the global model, the First/Second-Age City and the urban fabric composite models. Despite the set of goodness-of-fit measures previously described indicating higher performance values associated with the combined models, these differences were hardly detectable in the geographical space. An overall underprediction was observed for hilly neighborhoods surrounding the city center, and underprediction was also observed along the coastline, despite the inclusion of a specific set of indicators. Only expert-based knowledge of the study area might allow us to better understand and explain the underlying reasons for specific hyper-local over/under prediction values. For instance, underprediction was observed in correspondence with pedestrian areas or along those streets characterized by specific retail functional agglomeration issues resulting from historical/commercial inertia of the street/neighborhood [22] (p. 120).

4.2. Variable Selection

In this section, the results of the variable selection are presented. This second phase of the analysis allows us to identify and describe which combinations of indicators underlied the spatial distribution of micro-retail in the global study area as well as in each morphological region. A specific geographical/urban discussion of the individual roles of each urban form indicator goes beyond the goals of this paper. However, we provide some observations about the methodological procedures and an overall presentation of the selected variables.

In the global model as well as in the local models UF1, UF5 and UF6, the zero part was completely erased, resulting in an NB model. The reasons for this difference can most likely be traced to the model selection procedure of Enet algorithms based on the minimization of BIC as well as the higher penalizing factor for a larger number of regressors. These results might support the idea that an NB model is, in certain cases, a simpler and more efficient solution. On the contrary, the ZINB approach, despite being the most performant solution when the full model was studied, became too complex when a smaller number of variables was investigated.

The variable selection procedure allows the importance of a restricted number of variables between 27 (for compact urban regions) and 11/13 (for suburban and less dense urban fabrics) to be highlighted. From the initial 105 variables, 54 appeared in at least one model, with half of them found in at least three models. The left column of Table 5 enumerates the 27 most recurrent indicators in descending order of the number of models, while the right column provides the variable ranks when considering the importance of each variable assessed as the sum of the absolute increase/decrease of the odds ratios observed in every model.

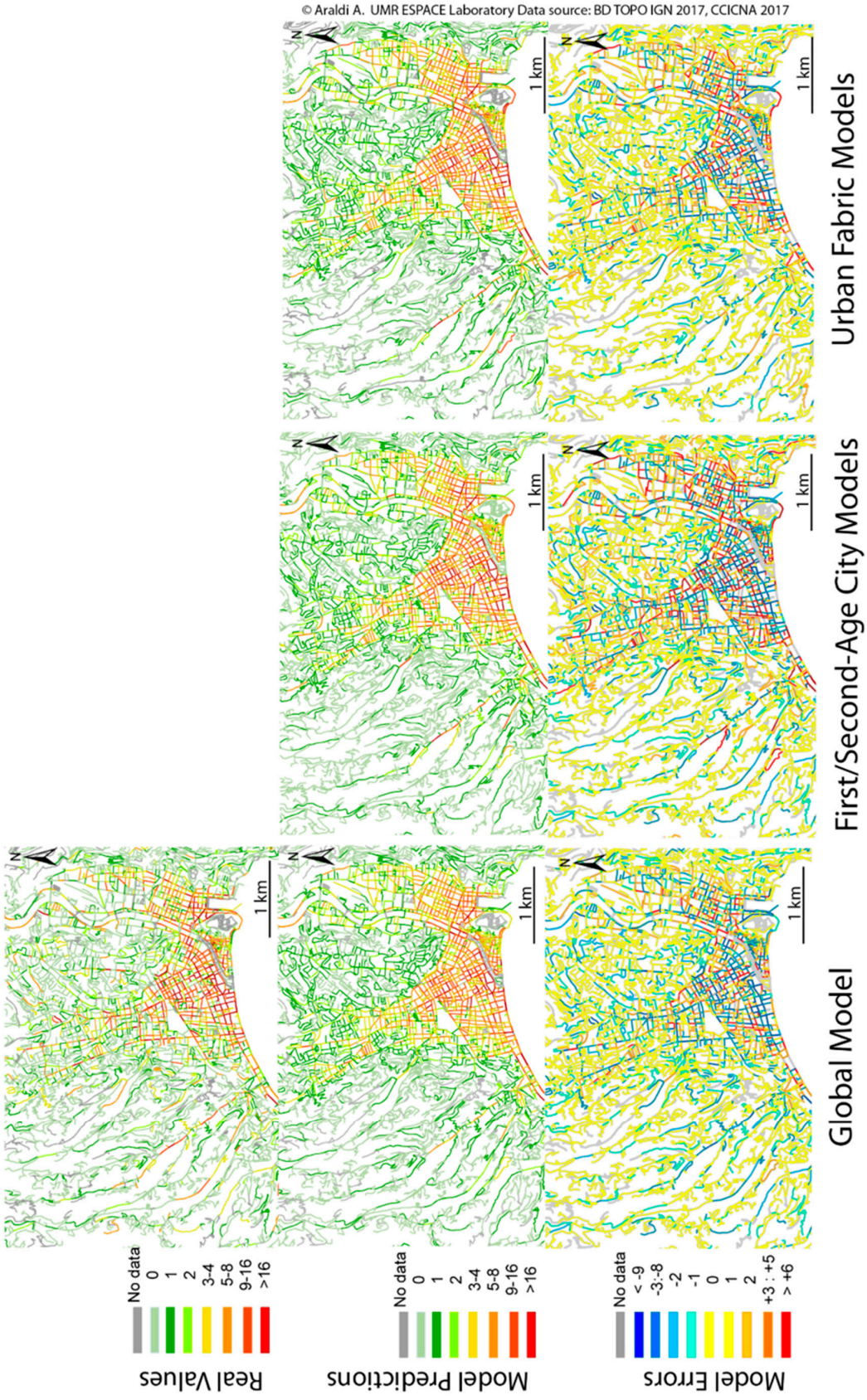

Figure 3. Projection in the geographical space of the observed, predicted and residual values for the global model, First/Second-Age City and urban fabric composite models.

Table 5. Outcomes of feature selection procedures. Selection frequencies of the most recurrent descriptors of urban form in relation to micro-retail spatial distribution

Indicator Ranking by			
N° Appearances		**Overall Impact**	
Betweenness 1200	9	Buil. Coverage Ratio	3.036
Street Acclivity	9	Betweenness 1200	2.087
Buil. Coverage Ratio	8	Street Acclivity	1.732
Street Corridor Effect	8	Buil. Fragmentation	1.364
Buil. Fragmentation	7	Street Corridor Effect	1.173
Avg. Build. Height	7	Street Length	1.121
Freq Parc	7	Avg Height	0.973
Avg. Open Space	6	Betweenness N 5	0.943
Between AS 1200	5	Avg. Street Wide	0.911
Street Length	4	Parcel Frequency	0.726
BetweennessN 5	4	UF7	0.563
StraightnessN 5	4	Avg SetBack	0.549
UF7	3	Std SetBack	0.504
Avg SetBack	3	Std Buil.Height	0.491
Std SetBack	3	UF4	0.483
UF4	3	Small Buil. (<125 m^2)	0.454
Small Buil. (<150 m^2)	3	Betw. Coast 600	0.445
Betw. Coast 600	3	Reach 20	0.433
Reach 20	3	Betweeness 600	0.381
Betweeness 600	3	Straightness coast	0.320
Straight. Coast 1200	3	Specialisation	0.317
Std. Open Space	3	Std. Open Space	0.315
Straightness 20	3	Straightness 20	0.300
StraightnessN 300	3	StraightnessN 5	0.264
Betw. Coast 2400	3	Reach 300	0.254
Straightness 1200	3	Closeness N 600	0.237
StraightnessN 1200 m	3	StraightnessN 1200 m	0.221

Note: Frequencies are here reported considering all nine models under analysis, ordered by number of appearances and overall impact (sum of the absolute increase/decrease of the odds ratios observed in all model). Background colors identify urban form descriptors categories: yellow, street-network configuration; light-green, skeletal streetscape, green, urban fabrics; blue, directional descriptors.

Table 6 presents all the selected indicators within each morphological region; variables selected for the count and zero parts are detailed in the upper and lower, respectively. Based on this table, we might observe how the built-up coverage ratio (PB50m), local betweenness (1200 m) and street acclivity represent the three aspects of urban form most related to the store distribution. This first outcome is in line with the results discussed in the urban form and micro-retail literature by [14], [8] and [87], respectively. The outcomes of this analysis show how micro-retail distribution might be explained by the combined effect of these three aspects (almost) independently according to the spatial partition under study (scale and contextual invariance). The built-up coverage ratio does not play a significant role in historical centers, UF1, where it has reached a certain homogeneity of high values (last phase of the burgage cycle [88]) and other urban form properties become more significant in defining favorable conditions for retail presence.

Table 6. Outcomes of the variable selection procedure (Enet-PR ZINB) implemented on the overall space of the French Riviera (global) and its contextual partitions (First-/Second-Age City, UF1–6).

COUNT-PART			Global	1st A.C.	2nd A.C.	UF1	UF2	UF3	UF4	UF5	UF6
	Impact	N° select	19	25	21	14	18	13	16	11	12
Built-up Coverage Ratio	3.036	8	2.005	1.234	1.753		1.152	1.308	1.213	1.368	1.002
Betw. 1200 m	2.023	7	1.160	1.169		1.031		1.486	1.704	1.361	1.112
Street Acclivity	1.732	9	0.854	0.699	0.842	0.759	0.730	0.878	0.725	0.963	0.819
Built-up Fragmentation	1.204	5	1.303	1.442			1.407		1.041	1.010	
Street Length	1.121	4	1.147	1.329	1.252				1.392		
Corridor Effect	1.061	4	1.212	1.281		1.399	1.169				
Avg. Height	0.973	7	1.155	1.072	1.136	1.376	1.009	1.197			1.028
Betw. N 5 m	0.943	4	1.232	1.126	1.393		1.193				
Avg. Open Space	0.911	6	1.127		1.076			1.042	1.262	1.024	1.379
Parcel Frequency	0.578	4			1.135			1.248		0.981	0.823
UF7	0.563	3	1.234	0.881	1.210						
Avg. Setback	0.549	3							1.297	1.116	1.136
Std. Setback	0.504	3		0.896			0.822		0.778		
Std. Height	0.491	2								1.132	1.359
UF4	0.483	3	1.169	0.915	1.229						
Small Build. (<125 m^2)	0.454	3		0.829	0.907		0.810				
Betw. Coast 600 m	0.445	3			1.096	1.075			1.275		
Reach 20 min	0.433	3	1.050		1.154	1.228					
Betw. 600 m	0.381	3	1.073			1.254	1.055				
Straig. Coast 1200 m	0.320	3	1.115	1.076			1.128				
Build. Specialization	0.317	1			1.317						
Std. Open Space	0.315	3		0.994		0.755	0.936				
Straig. 20 min	0.300	3	1.146	1.036			1.117				
Straig. N 5 min	0.264	4	1.039	1.091		1.094			1.039		
Reach 300	0.254	1				1.254					
Clos. N 600 m	0.237	2				0.896					0.868
Straig. N 300 m	0.221	3		1.115				0.912	1.018		
Reach N 5 min	0.205	2						0.859		1.064	
Betw. Coast 2400 m	0.199	3	1.037	1.093				1.069			
Betw. Coast 1200 m	0.191	2			1.038	1.153					
Betw. AS 1200 m	0.170	5	1.005	1.079	1.024	1.059			1.004		
Reach N 600 m	0.159	1									0.841
Straig. Places 300 m	0.153	2		1.050				0.897			
Reach Coast 1200 m	0.153	1				1.153					
Nodes 4	0.145	1					1.145				
Straig. 1200 m	0.139	3	1.034	1.075	1.030						
Reach 600 m	0.123	2						1.008			0.884
UF5	0.121	1		0.879							
Betw. 300 m	0.118	1								1.118	
Std. HW Ratio	0.115	1		0.885							
Straig. 5 min	0.115	1							1.115		
UF3	0.112	1		0.888							
Small Build. (125–250 m^2)	0.105	2			0.937		0.958				
Reach 1200 m	0.086	1						1.086			
Betw. Places 1200 m	0.080	1					1.080				
Straig. AS 600 m	0.079	2			1.002				1.077		
Straig.N 1200 m	0.067	3		1.014	1.025				1.029		
Straig. AS 1200 m	0.041	2			1.033				1.008		
AVG HW	0.041	1					0.959				
Large Build. (250–1000 m^2)	0.038	1			1.038						
Straig. N 600 m	0.035	2						1.029		1.006	
Betw. N 300 m	0.029	1									1.029
Betw. N 600 m	0.009	1					1.009				
Clos. 5 m	0.006	1					0.994				
ZERO-PART	Impact	N° select	0	2	2	0	4	2	2	0	0
Built-up Fragmentation	0.153	2		0.951			0.896				
Parcel Frequency	0.144	3			0.925		0.941		0.989		
Reach 5 min	0.13	1					0.870				
Corridor Effect	0.11	4		0.976	0.987		0.954	0.973			
Betw. 1200 m	0.062	2						0.994	0.943		

Note: Variables are ordered by impact factor; count and zero parts are separately described in the upper and lower parts, respectively. Background colors identify urban form descriptors categories: Yellow, street-network configuration; light-green, skeletal streetscape; green, urban fabrics; blue, directional descriptors.

Skeletal streetscape morphometric descriptors such as the built-up coverage ratio, the corridor effect, built-up fragmentation, average building height, open space, street acclivity and length have

a higher importance as locational factors in micro-retail distribution. These indicators are the most frequently selected, showing higher odd ratios compared to street-network configurational properties.

Indicators always negatively associated with micro-retail distribution are street acclivity, average and standard deviations of building setback and prevalence of small houses (footprint surface < 150 m^2).

The procedure implemented in this paper highlights the twofold role played by contextual descriptors. The first role is the direct influence of urban fabrics, morphological regions and their combinations on the definition of the retail presence; for example, both artificial connective and modernist fabrics (UF7 and UF4, respectively) were negatively correlated with micro-retail distribution when found within compact regions (First-Age City); on the contrary, they become positively associated with store distribution when located in car-oriented peripheral regions (Second-Age City). This observation supports the hypothesis of a double urban system that has been traditionally described by both urban form and micro-retail geographer researchers. The second role is the indirect effect on the variable selection procedure implemented within each region. While some streetscape and street-network configurational descriptors showed high values in every sub-region (i.e., built-up coverage ratio, local betweenness and street acclivity), others showed a high dependency on morphological context. In particular, some variables showed a significant role only in specific regions (i.e., corridor effect and building height were positively related to retail count only in compact fabrics, while the average set-back was negatively related to retail count only in suburban fabrics; Table 6), while others showed a divergent effect (i.e., parcel frequency showed positive/negative values for compact/open fabrics, respectively; Table 6). The identification of these regionalized behaviors would not have been possible with traditional global approaches. Moreover, these specific outcomes suggest the presence of more complex, non-linear relationships, with retail distribution requiring the exploration of more sophisticated modelling approaches.

When focusing on zero parts, we might notice how regression coefficients showed lower absolute values, and their impacts were always negatively related to an absence of micro-retail. Five indicators were selected: corridor effect, built-up fragmentation, parcel frequency, 5-min reach and 1200-m betweenness, each one utilized in different sub-regions.

5. Discussion and Conclusions

This paper presented and discussed some methodological aspects that researchers should consider when analyzing the relationship between micro-retail distribution and urban form from a street-based perspective.

In the first part of the paper, we highlighted how analytical approaches should account for the discrete, non-negative, highly skewed and zero-inflated nature of store distribution. Overlooking these aspects might affect modelling outcomes with both statistical and survivorship biases. Thus, identifying and implementing adapted modelling procedures becomes of paramount importance. Moreover, multicollinearity issues might arise from the assessment of a large number of urban form descriptors differently combined depending on their relative morphological contexts. Innovative modelling approaches are required to allow the evaluation of the combined effects of a large number of variables and to highlight their individual/relative contributions to an understanding of retail distribution. The final goal is to overcome the fragmented knowledge, providing a wider and holistic description of urban form and its relationship with micro-retail distribution.

To overcome these limitations, in the second part of this work we proposed the implementation of modelling and variable selection procedures within an integrated methodological framework. Seven count regression approaches were implemented (G, P, NB, ZIP, ZINB, ZAP, ZANB) in the real-world case study of the French Riviera metropolitan conurbation. The goal of these models was to estimate the number of stores per street segment from a dataset of 105 street-based descriptors of urban form (including street-network configurational properties, morphological skeletal streetscape and urban morphological contextual descriptors). A specific modelling selection procedure based on AIC and LR

tests allowed us to assess the performance levels of these seven models and highlight the superiority of the ZINB solution. The same conclusion was also reached when implementing the same model selection procedure separately in different morphological contexts defined at different scales. These outcomes confirm the hypothesis about the presence of a double-generating process at the origin of retail distribution that described the presence/absence and total number of stores observed along street segments.

Finally, the implementation of penalized regression procedures allowed us to select a reduced subset of urban form descriptors for each morphological region. Some indicators were significantly related to the retail distribution independent of the scale/context definition, while others assumed a specific role within given morphological subspaces. This outcome highlights the importance of the morphological context in the study of micro-retail distribution in metropolitan areas. This same outcome might also be interpreted from an urban planning and design perspective, as the need to study intrinsic properties of the urban form (i.e., streetscapes) depends on the general patterns/context within a multiscale/multilevel approach.

From an analytical perspective, this work provided a robust methodological framework for the study of retail distribution and urban form. Further works will examine the geographical and urban significance of these results as well as their contributions to the established theoretical framework of both urban form and retail geography.

The same methodological framework presented in this paper might also be implemented (with few adaptations) in the follow ways: (i) considering different/new urban form properties, with other functional and socioeconomic descriptors being included for a wider definition of the urban environment beyond the form of a physical city; (ii) considering specific retail categories, formats (i.e., franchise/independent stores) and surface-based categories; (iii) to assess the relative importance of urban descriptor categories (i.e., configurational, morphological and streetscape descriptors) or assess the capacity of different urban form protocols (i.e., SSx, MCA, etc.); (iv) for synchronic/diachronic comparative analysis; and (v) in conjunction with other human-based phenomena characterized by a discrete pattern of occurrences and measured on a fine-grained partition of the urban space.

Finally, future works could explore three main methodological aspects that are still overlooked in this work. Firstly, as regards the possible between-class variability of the hierarchical nested organization of the urban form (street, neighborhood, morphological regions), one solution might consider the implementation of MLM [44] combined with the GLM and PR procedures discussed in this work. Secondly, researchers might be interested in non-linear behaviors in data; indeed, the only downside of GLM procedures is the underlying hypothesis of a (generalized) linear relationship between the target variable and regressors. Machine learning modelling procedures should be tested both for modelling and variable selection procedures. Finally, a third aspect that could also be integrated with the four aspects discussed in this paper is the role of the spatial organization of stores; applying methodological approaches such as semivariograms and correlograms to both observed distribution and model errors [89].

Funding: This research was funded by Chambre du Commerce et d'Industrie Nice Côte d'Azur, CIFRE-EPACE Agreement N° 2015/1478.

Acknowledgments: The author of this paper would like to thank Giovanni Fusco, research fellow at ESPACE laboratory, for the valuable insights and the support in every phase of this research project.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike Information Criteria

BIC	Bayesian Information Criteria
MCA	Multiple Centrality Assessment
MFA	Multiple Fabric Assessment
Enet	Elastic Net Penalized Regression
GLM	Generalized Linear Model
MLR	Multiple Linear Regression
NB	Negative Binomial
P	Poisson
PR	Penalized Regression
SSx	Space Syntax
ZA	Zero Altered
ZANB	Zero Altered Negative Binomial
ZAP	Zero Altered Poisson
ZI	Zero Inflated
ZINB	Zero Inflated Negative Binomial
ZIP	Zero Inflated Poisson

Appendix A. Urban Form and Retail Distribution Literature Review

Table A1. Literature review: papers investigating the relationship between urban form (mainly street-network configuration properties) and micro-retail distribution. E/NC-KDE, Euclidean/network-constrained Kernel Density Estimation; CDF, cumulative distribution function; MLR/BLR, muliple/bivariate linear regression; ExpR, exponential regression; P-corr, Pearson correlation; NBR, negative binomial regression; K-W H-test, Kruskal–Wallis H test.

Urban Form and Micro-retail Distribution					
Authors	**Year**	**Dependent Variable**	**Phenomena**	**Space Study**	**Analytical Approach**
Hillier	1999	N° Stores/street	Micro-retail pattern	Camden London, UK	MLR
Cutini	2001	N° Stores/25 m (100 streets)	Micro-retail pattern	3 small-medim sized Italian towns	ExpR
Van Nes	2005	-	Micro-retail aggl.	Amsterdam, Netherland	Visual
Joosten and Van Nes	2005	-	Catering businesses	Berlin, Germany	Visual
Sarma	2006	N° Stores/aggl.	Micro-retail aggl.	New Delhi, India	BLR
Porta	2006	E-KDE: 100m-cells, bandwidth 100–300)	Micro-retail pattern	Bologna, Italy	P-Corr
Ortiz-Chao	2008	N° Stores/street	Micro-retail (land use)	Mexico City, Mexico	CDF
Porta	2012	E-KDE (300-mt BW) on a 10m-size cell raster	Micro-retail Pattern	Barcelona, Spain	P-Corr
Tsou, Chen	2013	Micro-retail density within traffic zones	Micro-retail Pattern	Taipei city, taiwan	MlogLR
Van Nes	2014	-	Micro-retail Pattern	Pompeii, Rome	Visual
Wang et al.	2014	E-KDE (1.5-km BW on 100-m cell-side raster)	Micro-retail pattern	ChangChun, China	P-corr
Sevtsuk Sevtsuk	2014 2010	Presence/absence micro-retail building level	Micro-retail pattern	Cambridge and Sommerville, USA	MLR-Spatial Lag and Error
Cui and Han	2015	E-KDE (1.5-km BW, 100-m size cell)	Micro-retail (Point of Interest)	Zhengzhou, China	P-corr
Omer and Goldblatt	2015	N° Build. with micro-retail 50m street-buffer	Micro-retail pattern (comm.build.)	8 Israeli Cities (3 types)	P-corr MLR
Scoppa Peponis Scoppa	2013 2015	Micro-retail frontage/ street length	Micro-retail pattern (comm.parcels)	Buenos Aires, Argentina	BLR, PCA-MLR
Ye et al.	2017	N° Stores/street block	Catering businesses	Shenzhen, China	NBR
Lin et al.	2018	E-KDE (3.5-km BW, 100-m size cells)	Micro-retail pattern (POI)	Guangzhou, China	
Cutini et al.	2018	N° Stores/ street (30 streets)	Micro-retail pattern	Milan, Italy	Exp-Corr
Saraiva et al.	2019	E-KDE (20-m size cells)	Micro-retail vacancy	4 medium-sized Portuguese cities	P-corr
Bobkova et al.	2019	N° Stores/ plot	Micro retail pattern	London, Amsterdam Stockholm	K-W H-test

Appendix B. Street-based Urban Form Measures

Appendix B.1. Street Network Configurational Indicators

Following the definition by Porta et al. [8,9] of street network centrality indicators, for each street midpoint i lying on the network G, we implement:

$$Reach_r(i) = \sum_{j \in G-\{i\}; d[i,j] \leq r} j \tag{A1}$$

$$Closeness_r(i) = \frac{1}{\sum_{j \in G-\{i\}; d[i,j] \leq r} d[i,j]} \tag{A2}$$

$$Straighntess_r(i) = \sum_{j \in G-\{i\}; d[i,j] \leq r} \frac{\delta[i,j]}{d[i,j]} \tag{A3}$$

$$Betweenness_r(i) = \sum_{j,k \in G-\{i\}; d[j,k] \leq r} n_{jk}[i] \tag{A4}$$

where:

1. $d[i, j]$ represents the distance of the shortest path between the reference midpoint i and each destination midpoint j within the sub-network identified by the radius r;
2. $\delta[i, j]$ represents the relative Euclidean distance between each midpoint i and each destination midpoint j within the same distance;
3. $n_{jk}[i]$ is the number of minimum paths from node j to node k on network G passing through point i, with j and k at a distance less than or equal to r.

Following the definition by Luo and Wang [63] of a two-step floating catchment area (2SFCA), we implement the normalization (N) of Equations (A1)–(A4) as:

$$Reach\ _r^N(i) = \sum_{j \in G-\{i\}; d[i,j] \leq r} \left(\frac{j}{R_r(j)} \right) \tag{A5}$$

$$Closeness_r^N(i) = \frac{1}{\sum_{j \in G-\{i\}; d[i,j] \leq r} d[i,j] \cdot \frac{1}{R_r(j)}} \tag{A6}$$

$$Straightness_r^N(i) = \sum_{j \in G-\{i\}; d[i,j] \leq r} \frac{\delta[i,j]}{d[i,j]} \cdot \frac{1}{R_r(j)} \tag{A7}$$

$$Betweenness_r^N(i) = \sum_{j,k \in G-\{i\}; d[j,k] \leq r} n_{jk}[i] \cdot \frac{1}{R_r(j)} \tag{A8}$$

where $R_r(j)$ is the $Reach_r$ of each street midpoint j within the sub-network identified by the radius r (as defined in Equation (A1)).

Table A2. Summary table of the 40 street-network configurational indicators. r, radius; n, normalized.

	Pedestrian r [meters]			**Vehicle [minutes]**	
	300	**600**	**1200**	**5**	**20**
$Reach\ _r$	R_{300}	R_{600}	R_{1200}	R_5	R_{20}
$Reach\ _r^N$	R_{300}^N	R_{600}^N	R_{1200}^N	R_5^N	R_{20}^N
$Closeness_r$	C_{300}	C_{600}	C_{1200}	C_5	C_{20}
$Closeness_r^N$	C_{300}^N	C_{600}^N	C_{1200}^N	C_5^N	C_{20}^N
$Straightness_r$	S_{300}	S_{600}	S_{1200}	S_5	S_{20}
$Straightness_r^N$	S_{300}^N	S_{600}^N	S_{1200}^N	S_5^N	S_{20}^N
$Betweenness_r$	B_{300}	B_{600}	B_{1200}	B_5	B_{20}
$Betweenness_r^N$	B_{300}^N	B_{600}^N	B_{1200}^N	B_5^N	B_{20}^N

Equations (A1)–(A4) are implemented considering a specific weight matrix associating each midpoint *j* with the presence/absence of a given urban features (such as squares, coastline and anchor stores). This approach would allow the directional centrality measures summarized in the following table to be obtained.

Table A3. Summary table of the 36 directional centrality indicators. *r*, radius; *S*, squares; *C*, coastline, *AS*, anchor stores.

	Towards Squares *r* [metres]			**Towards Coastline *r* [meters]**			**Towards Anchor Stores *r* [meters]**		
	300	**600**	**1200**	**600**	**1200**	**2400**	**300**	**600**	**1200**
$Reach_r$	R^S_{300}	R^S_{600}	R^S_{1200}	R^C_{300}	R^C_{600}	R^C_{1200}	R^{AS}_{300}	R^{AS}_{600}	R^{AS}_{1200}
$Closeness_r$	C^S_{300}	C^S_{600}	C^S_{1200}	C^C_{300}	C^C_{600}	C^C_{1200}	C^{AS}_{300}	C^{AS}_{600}	C^{AS}_{1200}
$Straightness_r$	S^S_{300}	S^S_{600}	S^S_{1200}	S^C_{300}	S^C_{600}	S^C_{1200}	S^{AS}_{300}	S^{AS}_{600}	S^{AS}_{1200}
$Betweenness_r$	B^S_{300}	B^S_{600}	B^S_{1200}	B^C_{300}	B^C_{600}	B^C_{1200}	B^{AS}_{300}	B^{AS}_{600}	B^{AS}_{1200}

Appendix B.2. Skeletal Streetscape Descriptors

Figure A1 proposes a schematic illustration of the two GIS protocols implemented for the description of the skeletal streetscape. While the sightline approach (on the left) describes the façade disposition along the street centerline (Table A4), the proximity band approach (on the right) allows the description of the building masses surface/volumetric distribution (Table A5).

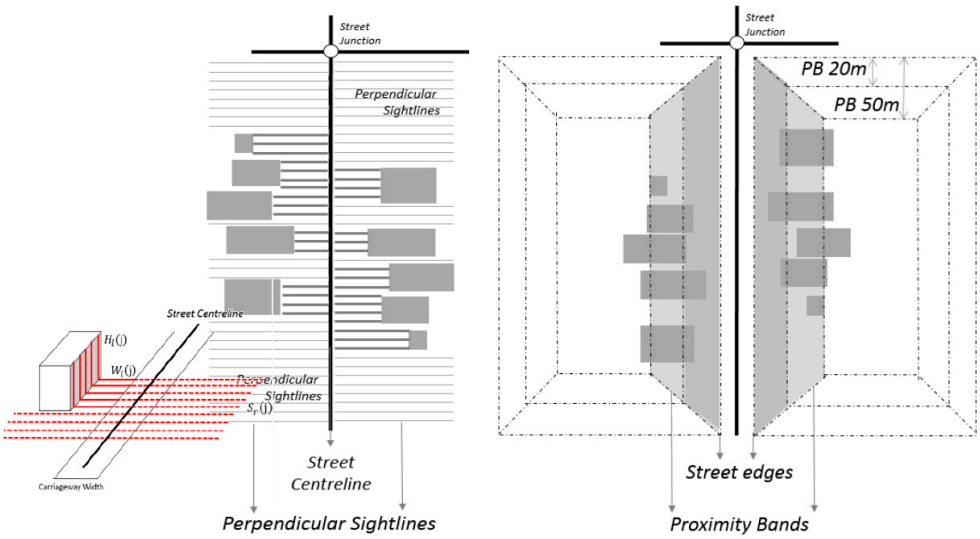

Figure A1. Graphical representation of the two skeletal streetscape GIS protocols. On the left: building façade described through sightlines perpendicular to the street centerline, homogeneously distributed (3 m). On the right: building footprint and volumes captured by the proximity band approach (source: [89]).

Table A4. Summary table of the 36 directional centrality indicators. *r*, radius; *S*, squares; *C*, coastline; *AS*, anchor stores.

Streetscape Indicator from Street Sightlines		
Urban Streetscape Component	**Indicator**	**Implementation Formulae**
Open Space	*Openness*	$\frac{1}{N}\sum_{j=1}^{N} S_r(j) + S_l(j)$
	Openness Roughness	$\sqrt{\frac{\left(\sum_{j=1}^{N}\left(S_r(j)-\overline{S_r(j)}\right)+\sum_{j=1}^{N}\left(S_l(j)-\overline{S_l(j)}\right)\right)}{N-1}}$

Table A4. *Cont.*

Streetscape Indicator from Street Sightlines		
Urban Streetscape Component	**Indicator**	**Implementation Formulae**
Facades-Street Network-Parcels Relationship	*Building Setback* *	$\frac{1}{n}\sum_{j=1}^{n} W_r(\mathrm{j}) + W_l(\mathrm{j})$
	Facades Misalignment	$\sqrt{\frac{\left(\sum_{j=1}^{n}\left(W_r(\mathrm{j})-\overline{W_r(\mathrm{j})}\right)^2\right)}{n_r-1} + \frac{\left(\sum_{j=1}^{n}\left(W_l(\mathrm{j})-\overline{W_l(\mathrm{j})}\right)^2\right)}{n_l-1}}$
	Average Building Height	$\frac{1}{n}\sum_{j=1}^{n} H_r(\mathrm{j}) + H_l(\mathrm{j})$
	Building Height Misalignment	$\sqrt{\frac{\left(\sum_{j=1}^{n}\left(H_r(\mathrm{j})-\overline{H_r(\mathrm{j})}\right)^2\right)}{n_r-1} + \frac{\left(\sum_{j=1}^{n}\left(H_l(\mathrm{j})-\overline{H_l(\mathrm{j})}\right)^2\right)}{n_l-1}}$
Facades Cross-sectional Ratio	*Cross-sectional proportion*	$\frac{1}{n}\sum_{j=1}^{n} HW_r(\mathrm{j}) + HW_l(\mathrm{j})$
	Variability of Cross-sectional proportion	$\sqrt{\frac{\left(\sum_{j=1}^{n}\left(HW_r(\mathrm{j})-\overline{HW_r(\mathrm{j})}\right)^2\right)}{n_r-1} + \frac{\left(\sum_{j=1}^{n}\left(HW_l(\mathrm{j})-\overline{HW_l(\mathrm{j})}\right)^2\right)}{n_l-1}}$

Table A5. Streetscape indicators implemented through the proximity band procedure (source: [56]).

Streetscape Indicator from Proximity Bands				
Urban Fabric Component	***Indicator***	**Definition and Implementation Formulae**		**Proximity Band Width**
Network Morphology	*Street Length*	Street segments length between two intersections	L_{street}	/
	Windingness	1−(Euclidean distance/network distance) between two intersections	$1 - \frac{L_{eucl.}}{L_{street}}$	/
	Local connectivity	Average of the presence nodes of degree 1 (ND1)	$\sum \mathrm{ND}_i[0,1]/2$	/
		Average presence nodes of degree 4 (ND4)		/
		Average presence nodes of degree 3, 5 + (ND35+)		/
Built-up Morphology	*Prevalence of Building types*	(0:125] m2 building surf./total built-up surf.	$\frac{\sum Sj}{S_{builted}}$	50
		(125:250] m2 building surf./total built-up surf.		
		(250:1000] m2 building surf./total built-up surf.		
		(1000:4000] m2 building surf./total built-up surf.		
		(4000: max] m2 building surf./total built-up surf.		
	PB coverage ratio	Built-up Surface/PB Surf.	$\sum S_{tot} / \sum S_{PB}$	
	Building Contiguity	Weighted average of buildings frequency on built-up units	$\frac{\sum S_{b-u(i)}\left(\frac{1}{N_{build\ in\ b-u(i)}}\right)}{\sum S_{b-u\ (i)}}$	
	Specialization of Building Types	Specialized Building surf./PB surf.	$\frac{\sum S_{spec}}{\sum S_{PB}}$	
Network-Building Relationship	*Street corridor effect*	Parallel façades length/street length	$L_{par.fac}/L_{street}$	10
	PB building height H	Building volume/PB surface	$\sum V_{builted} / \sum S_{builted}$	20
	Open Space Width W	(PB surf.-built surf.)/street length	$(S_{PB}-S_{built}\ )/L_{street}$	
	Height/Width Ratio	PB Building Height/Open Space Width	H/W	
	Building frequency along SN	N. of Buildings/Str. length	N_{build}/L_{street}	
Network-Plot Relationship	*Parcel Frequency*	N. of Plots/Street length	N_{plot}/L_{street}	50
Site Morphology	*Surface slope*	High sloped surf. (S > 30%)/PB Surface	$\sum Sloped\ Surf_i / S_{PB}$	50
Network-Site Relationship	*Street acclivity*	Avg. arct(slope) along the street centerline	$E\ [arct(slope)_i]$	/

Appendix B.3. Urban Fabrics

Figure A2. Aerial and street view images of the nine urban fabrics (UFs) of the French Riviera as defined by the MFA protocol (source: Google Map and Google Street view 2017, [56,57,89]).

References

1. Smith, A.; Sparks, L. The role and function of the independent small shop: The situation in Scotland. *Int. Rev. Retail Distrib. Consum. Res.* **2000**, *10*, 205–226. [CrossRef]
2. Chiaradia, A.; Hillier, B.; Schwander, C.; Wedderburn, M. Spatial Centrality, Economic Vitality/Viability. In *Proceedings of the 7th International Space Syntax Symposium*; KTH Royal Institute of Technology: Stockholm, Sweden, 2009.
3. Aversa, J.; Doherty, S.; Hernandez, T. Big Data Analytics: The New Boundaries of Retail Location Decision Making. *Pap. Appl. Geogr.* **2018**, *4*, 390–408. [CrossRef]
4. Saraiva, M.M. The Morphological Sense of Commerce: Symbioses between Commercial Activity and the Form and Structure of Portuguese Medium-Sized Cities. Ph.D. Thesis, Univ. do Porto, Porto, Portugal, 2013.
5. Hillier, B. *Space is the Machine*; Cambridge University Press: Cambridge, UK, 1996.
6. Hillier, B.; Iida, S. Network and Psychological Effects in Urban Movement. In *Spatial Information Theory*; Cohn, A., Mark, D., Eds.; Springer: Berlin, Germany, 2005; pp. 475–490.
7. Hillier, B. Centrality as a process. *Urban Des. Int.* **1999**, *4*, 107–127. [CrossRef]
8. Porta, S.; Strano, E.; Iacoviello, V.; Messora, R.; Latora, V.; Cardillo, A.; Scellato, S. Street centrality and densities of retail and services in Bologna, Italy. *Environ. Plan. B Plan. Des.* **2009**, *36*, 450–465. [CrossRef]
9. Porta, S.; Latora, V.; Wang, F.; Rueda, S.; Strano, E.; Scellato, S.; Latora, L. Street centrality and the location of economic activities in Barcelona. *Urban Stud.* **2012**, *49*, 1471–1488. [CrossRef]
10. Saraiva, M.; Marques, T.S.; Pinho, P. Vacant Shops in a Crisis Period—A Morphological Analysis in Portuguese Medium-Sized Cities. *Plan. Pract. Res.* **2019**, *34*, 255–287. [CrossRef]
11. Remali, A.M.; Porta, S.; Romice, O. Correlating street quality, street life and street centrality in Tripoli, Libya. 2014. Available online: https://strathprints.strath.ac.uk/50265/ (accessed on 23 March 2020).
12. Ye, Y.; Li, D.; Liu, X. How block density and typology affect urban vitality: An exploratory analysis in Shenzhen, China. *Urban Geogr.* **2018**, *39*, 631–652. [CrossRef]
13. Joosten, V.; Van Nes, A. How block types influence the natural movement economic process: Micro-spatial conditions on the dispersal of shops and Café in Berlin. In Proceedings of the 5th International Space Syntax Symposium, Delft, The Netherlands, 13–17 June 2005; Volume 13.
14. Bobkova, E.; Marcus, L.; Berghauser Pont, M.; Stavroulaki, I.; Bolin, D. Structure of plot systems and economic activity in cities: Linking plot types to retail and food services in London, Amsterdam and Stockholm. *Urban Sci.* **2019**, *3*, 66. [CrossRef]
15. Saraiva, M.; Pinho, P. Spatial modelling of commercial spaces in medium-sized cities. *GeoJournal* **2017**, *82*, 433–454. [CrossRef]
16. Cutini, V. Centrality and land use: Three case studies on the configurational hypothesis. *Cybergeo* **2001**, *10*. [CrossRef]
17. Omer, I.; Goldblatt, R. Spatial patterns of retail activity and street network structure in new and traditional Israeli cities. *Urban Geogr.* **2016**, *37*, 629–649. [CrossRef]
18. Wang, F.; Antipova, A.; Porta, S. Street centrality and land use intensity in Baton Rouge, Louisiana. *J. Transp. Geogr.* **2011**, *19*, 285–293. [CrossRef]
19. Cui, C.; Han, Z. Spatial patterns of retail stores using POIs data in Zhengzhou, China. In Proceedings of the 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), Fuzhou, China, 8–10 June 2015.
20. Wang, S.; Xu, G.; Guo, Q. Street centralities and land use intensities based on points of interest (POI) in Shenzhen, China. *Int. J. Geo-Inf.* **2018**, *7*, 425. [CrossRef]
21. Cutini, V.; Farese, D.; Rabino, G. Milan: The Configuration of a Metropolis. In *Smart Planning: Sustainability and Mobility in the Age of Change*; Springer: Cham, Switzerland, 2018; pp. 343–357.
22. Sevtsuk, A. Path and Place: A Study of Urban Geometry and Retail Activity in Cambridge and Somerville. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2010.
23. Sevtsuk, A. Location and agglomeration: The distribution of retail and food businesses in dense urban environments. *J. Plan. Educ. Res.* **2014**, *34*, 374–393. [CrossRef]
24. Scoppa, M.D. Towards a Theory of Distributed Attraction: The Effects of Street Network Configuration Upon the Distribution of Retail in the City of Buenos Aires. Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2013.

25. Scoppa, M.D.; Peponis, J. Distributed attraction: The effects of street network connectivity upon the distribution of retail frontage in the City of Buenos Aires. *Environ. Plan. B Plan. Des.* **2015**, *42*, 354–378. [CrossRef]
26. Gardner, W.; Mulvey, E.P.; Shaw, E.C. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychol. Bull.* **1995**, *118*, 392. [CrossRef]
27. Pipkin, J.S. A Partitioning Model of Urban Retail Structure. *Geogr. Anal.* **1993**, *25*, 179–198. [CrossRef]
28. Lebrun, N. Centralités Urbaines et Concentrations de Commerces. Ph.D. Thesis, Université de Reims-Champagne Ardenne, Reims, France, 2002.
29. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*; Cambridge Univ. Press: Cambridge, UK, 2013; Volume 53.
30. Hilbe, J.M. *Negative Binomial Regression*; Cambridge University Press: Cambridge, UK, 2011.
31. Guy, C.M. Recent advances in spatial interaction modelling: An application to the forecasting of shopping travel. *Environ. Plan. A* **1987**, *19*, 173–186. [CrossRef]
32. Shonkwiler, J.S.; Harris, T.R. A Non-Gaussian Time Series Analysis of Rural Retail Business Counts. *J. Reg. Sci.* **1993**, *33*, 37–48. [CrossRef]
33. Taleb, N.N. *The Black Swan: The Impact of the Highly Improbable*; Random House: New York, NY, USA, 2007; Volume 2.
34. Heywood, I. *Introduction to Geographical Information Systems*; Addison Wesley Longman: New York, NY, USA, 1998.
35. Zhang, M.; Kukadia, N. Metrics of urban form and the modifiable areal unit problem. *Transp. Res. Rec.* **2005**, *1902*, 71–79. [CrossRef]
36. Holt, D.; Steel, D.G.; Tranmer, M.; Wrigley, N. Aggregation and ecological effects in geographically based data. *Geogr. Anal.* **1996**, *28*, 244–261. [CrossRef]
37. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **1992**, *34*, 1–14. [CrossRef]
38. Oakes, J.M.; Andrade, K.E.; Biyoow, I.M.; Cowan, L.T. Twenty years of neighborhood effect research: An assessment. *Curr. Epidemiol. Rep.* **2015**, *2*, 80–87. [CrossRef] [PubMed]
39. Jenks, C.; Mayer, S. The Consequences of Growing up in a Poor Neighborhood. In *Inner City Poverty in the United States*; Lynn, M., Ed.; McGeary: Washington, DC, USA, 1990; pp. 111–186.
40. Kaufman, J.S.; Cooper, R.S. Seeking causal explanations in social epidemiology. *Am. J. Epidemiol.* **1999**, *150*, 113–120. [CrossRef]
41. Greenland, S.; Morgenstern, H. Confounding in health research. *Annu. Rev. Public Health* **2001**, *22*, 189–212. [CrossRef]
42. Robinson, W.S. Ecological correlations and the behavior of individuals. *Am. Sociol. Rev.* **1950**, *15*, 351–357. [CrossRef]
43. Nezlek, J.B. *Multilevel Modeling for Social and Personality Psychology*, 1st ed.; SAGE: London, UK, 2011.
44. Kropf, K. Bridging configurational and urban tissue analysis. In Proceedings of the 11th Space Syntax Symposium, Lisbon, Portugal, 3–7 July 2017.
45. Kutner, M.H.; Nachtsheim, C.J.; Neter, J.; Li, W. *Applied Linear Statistical Models*; McGraw-Hill Irwin: Boston, MA, USA, 2005; Volume 5.
46. Cohen, J.; Cohen, P.; West, S.G.; Aiken, L.S. *Applied Multiple Correlation/Regression Analysis for the Social Sciences*, 3rd ed.; Erlbaum: Hillsboro, NJ, USA, 2003.
47. Craney, T.A.; Surles, J.G. Model-dependent variance inflation factor cutoff values. *Qual. Eng.* **2002**, *14*, 391–403. [CrossRef]
48. Judd, C.M.; McClelland, G.H.; Ryan, C.S. *Data Analysis: A Model Comparison Approach*; Harcourt Brace Jovanovich: San Diego, CA, USA, 2011.
49. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
50. Maddala, G.S.; Lahiri, K. *Introduction to Econometrics*, 4th ed.; Wiley: Hoboken, NJ, USA, 2009.
51. Lee, C.; Moudon, A.V. The 3Ds+ R: Quantifying land use and urban form correlates of walking. *Transp. Res. Part D Transp. Environ.* **2006**, *11*, 204–215. [CrossRef]
52. Wei, H.L.; Billings, S.A. Feature subset selection and ranking for data dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 162–166. [CrossRef]

53. Roth Tran, B. Blame It on the Rain: Weather Shocks and Retail Sales. 2016. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3381302 (accessed on 23 March 2020).
54. Vakhutinsky, A.; Mihic, K.; Wu, S.M. A Prescriptive Analytics Approach to Markdown Pricing for a in E-Commerce Retailer. *J. Pattern Recognit. Res.* **2019**, *1*, 1–21.
55. Verstraete, G.; Aghezzaf, E.H.; Desmet, B. A data-driven framework for predicting weather impact on high-volume low-margin retail products. *J. Retail. Consum. Serv.* **2019**, *48*, 169–177. [CrossRef]
56. Araldi, A.; Fusco, G. From the built environment along the street to the metropolitan region. Human scale approach in urban fabric analysis. *Environ. Plan B Urban Anal. City Sci.* **2019**, *46*, 1243–1263. [CrossRef]
57. Fusco, G.; Araldi, A. The Nine Forms of the French Riviera: Classifying Urban Fabrics from the Pedestrian Perspective. In *24th ISUF International Conference. Book of Papers (1313–1325)*; Editorial Universitat Politècnica de València: Valencia, Spain, 2017.
58. Ortiz-Chao, C.G. Land use patterns and access in Mexico City. In Proceedings of the ACSP-AESOP Fourth Joint Congress, Chicago, IL, USA, 6–11 July 2008.
59. Wang, Z. Regularized Linear Models. 2020. Available online: https://cran.r-project.org/web/packages/mpath/index.html (accessed on 23 March 2020).
60. Fleury, A. La rue: Un objet géographique? *Tracés. Revue Sci. Hum.* **2004**, *5*, 33–44. [CrossRef]
61. Marshall, S.; Gil, J.; Kropf, K.; Tomko, M.; Figueiredo, L. Street network studies: From networks to models and their representations. *Netw. Spat. Econ.* **2018**, *18*, 735–749. [CrossRef]
62. Batty, M. Agents, cells, and cities: New representational models for simulating multiscale urban dynamics. *Environ. Plan. A* **2005**, *37*, 1373–1394. [CrossRef]
63. Luo, W.; Wang, F. Spatial accessibility to primary care and physician shortage area designation: A case study in Illinois with GIS approaches. In *Geographic Information Systems and Health Applications*; Skinner, R., Khan, O., Eds.; Idea Group Publishing: Hershey, PA, USA, 2003; pp. 260–278.
64. Harvey, C.; Aultman-Hall, L.; Troy, A.; Hurley, S.E. Streetscape skeleton measurement and classification. *Environ. Plan. B Urban Anal. City Sci.* **2017**, *44*, 668–692. [CrossRef]
65. Purciel, M.; Neckerman, K.M.; Lovasi, G.S.; Quinn, J.W.; Weiss, C.; Bader, M.D.; Rundle, A. Creating and validating GIS measures of urban design for health research. *J. Environ. Psychol.* **2009**, *29*, 457–466. [CrossRef]
66. Vialard, A.A. Typology of Block-Faces. Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2013.
67. Araldi, A.; Perez, J.; Fusco, G.; Fuse, T. Multiple Fabric Assessment: Focus on Method Versatility and Flexibility. In *Computational Science and Its Applications—ICCSA2018. Proceedings, Part III, Lecture Notes in Computer Science*; Springer: Berlin, Germany, 2018; Volume 10962, pp. 251–267.
68. Portzamparc, C. *L'âge III*; Projet urbain, n° 3; La ville Hors la Ville: Paris, UK; Minist. Équipement: Paris, France, 1995; pp. 4–6.
69. McCullagh, P. *Generalized Linear Models*, 2nd ed.; Chapman and Hall/CRC: Horsham, PA, USA, 2018.
70. Agresti, A.; Kateri, M. *Categorical Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2011.
71. St-Pierre, A.P.; Shikon, V.; Schneider, D.C. Count data in biology—Data transformation or model reformation? *Ecol. Evol.* **2018**, *8*, 3077–3085. [CrossRef]
72. Kutner, M.H.; Neter, J.; Nachtsheim, C.J.; Li, W. *Applied Linear Regression Models*, 4th ed.; McGraw-Hill: New York, NY, USA, 2004.
73. Mullahy, J. Specification and testing of some modified count data models. *J. Econom.* **1986**, *33*, 341–365. [CrossRef]
74. King, G. Variance specification in event count models: From restrictive assumptions to a generalized estimator. *Am. J. Political Sci.* **1989**, *33*, 762–784. [CrossRef]
75. Cameron, A.C.; Windmeijer, F.A. An R-squared measure of goodness of fit for some common nonlinear regression models. *J. Econom.* **1997**, *77*, 329–342. [CrossRef]
76. Long, J.S.; Freese, J. *Regression Models for Categorical Dependent Variables Using Stata*; Revised Edition; Stata Press: College Station, TX, USA, 2003.
77. Mittlböck, M.; Schemper, M. Explained variation for logistic regression. *Stat. Med.* **1996**, *15*, 1987–1997. [CrossRef]
78. Menard, S. Coefficients of determination for multiple logistic regression analysis. *Am. Stat.* **2000**, *54*, 17–24.
79. Tjur, T. Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *Am. Stat.* **2009**, *63*, 366–372. [CrossRef]

80. Akaike, H. A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*; Springer: New York, NY, USA, 1974; pp. 215–222.
81. Vuong, Q.H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econom. J. Econom. Soc.* **1989**, *57*, 307–333. [CrossRef]
82. Wilson, P. The misuse of the Vuong test for non-nested models to test for zeroinflation. *Econ. Lett.* **2015**, *127*, 51–53. [CrossRef]
83. Kleiber, C.; Zeileis, A. Visualizing count data regressions using rootograms. *Am. Stat.* **2016**, *70*, 296–303. [CrossRef]
84. Gujarati, D.N. *Basic Econometrics*, 5th ed.; Tata McGraw-Hill Education Private Ltd.: New Delhi, India, 2009.
85. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; Volume 112, pp. 3–7.
86. Gideon, S. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
87. Parkins, A.E. Profiles of the Retail Business Section of Nashville, Tenn., and Their Interpretation. *Ann. Assoc. Am. Geogr.* **1930**, *20*, 164–175. [CrossRef]
88. Conzen, M.R.G. Alnwick, Northumberland: A study in town-plan analysis. *Trans. Pap. Inst. Br. Geogr.* **1960**, *27*, iii-122. [CrossRef]
89. Araldi, A. Retail Distribution and Urban Form: Street-Based Models for the French Riviera. Ph.D. Thesis, Université Côte d'Azur, Nice, France, 2019.