

IFN Working Paper No. 1558, 2026

Defining Geographic Markets Through Choice Sets: An Empirical Method Applied to Real Estate Brokerage

Adam Lindhe and Johan Orrenius

Defining Geographic Markets Through Choice Sets: An Empirical Method Applied to Real Estate Brokerage

Adam Lindhe & Johan Orrenius *

April 2026

Abstract

We introduce a choice-set approach to defining markets and a novel method to empirically recover geographic markets using machine learning, Spatial and Categorical Bayesian Clustering (SCBC). SCBC leverages the identity of the seller for each observation to capture market structures in a novel way that is not captured by purely distance-based methods. Applied to real estate agents in Stockholm (Sweden), SCBC classifies sales more accurately than the baseline K -means algorithm. Finally, we investigate the correct number of clusters and find that the optimal number of clusters is close to the validation set based on industry knowledge.

*Orrenius: Research Institute of Industrial Economics (IFN). Email johan.orrenius@ifn.se (corresponding author). Lindhe: KTH. We thank Petter Berg for generously providing data access. We thank Richard Friberg, Alon Eizenberg, Christos Genakos, Frode Steen, Henrik Horn, and Salil Sharma for their insightful comments and seminar participants at the Swedish Competition Authority 2025, IFN 2025, EARIE 2025. Johan Orrenius thanks Jan Wallanders och Tom Hedelius stiftelse for financial support

1 Introduction

Economic activity takes place in markets, but market boundaries are often hard to define. In economics, the logic is straightforward: a market consists of goods or services that can easily be substituted. Substitution has both a product and a geographic dimension. Products must compete for the same demand, and consumers must be able to switch across sellers within a relevant geographic area. This principle is central to competition analysis, but empirical market definition often remains weakly tied to it and instead relies on crude proxies.

For the geographical dimension, the contrast between principle and practice is stark. In principle, geographic markets should reflect the alternatives consumers can realistically consider. In practice, geographical markets are delineated using administrative boundaries, fixed radii, or travel-time cutoffs. Such proxies, although practical, imperfectly capture the substitution choices available to consumers. Factors shaping the geographic scope of substitution introduced in the U.S. Department of Justice and Federal Trade Commission (2023) new merger guidelines, such as natural barriers, local reputation, and institutional constraints, might be overlooked.

This article first proposes a choice-set framework for market definition. We define a market as a set of consumers facing the same, or sufficiently similar, feasible alternatives. To avoid requiring each consumer to be a separate market, we use a probabilistic notion of market membership: consumers belong to the same market when their available set of sellers are similar enough in an economically meaningful way. In geographic settings, this definition treats markets as a set of consumers who share a common choice environment rather than arbitrary pre-drawn areas or crude grouping by distance alone. The theoretical contribution is to formalize a market definition that connects directly to substitution opportunities as represented by consumers' choice sets without considering de-

mand or elasticities as previous theoretically based methods have done. The advantage of the choice sets approach is that when implemented it is tractable and requires minimal data.

Secondly, the article takes the choice set framework to data. Our novel empirical method, Spatial and Categorical Bayesian Clustering (SCBC), recovers latent geographic markets from observed transactions using only consumer location and seller identity. The key idea is to combine spatial proximity with similarity in consumer-seller links, and thereby recover shared choice environments that are influenced by factors other than distance. SCBC is an estimation strategy, not the market definition itself. We apply the method to real-estate agents in Stockholm as a motivating example. SCBC then recovers market boundaries more accurately than a purely distance-based benchmark.

Our article’s main contribution to the literature is on the definition of geographic markets. We introduce a novel theory-based definition of markets grounded in consumers’ choice sets and link it to a straightforward estimation method for delineating markets in settings where consumer locations and seller identity are observed. This gives us an applicable method with a strong link to the concept of substitution.

As Elzinga & Howell (2018) notes, geographic market definition has historically followed two approaches: demand-based methods using prices, market power, or cost functions, and proxy-based methods using distance or administrative units. Demand-based methods such as the hypothetical monopolist test require detailed data, natural experiments, or restrictive assumptions (Stigler, G ; Sherwin, 1985; Scheffman & Spiller, 1987). Our approach is empirically closer to the second category, requiring little data and few assumptions, while also incorporating substitution logic.

The choice of market definition matters as it qualitatively changes the mea-

sures of competition and the conclusions we draw from them, as seen in e.g. Elzinga & Howell (2018) and Genakos & Pagliero (2022). The Poundland-99p merger case (Competition and Markets Authority, 2015) is a prominent example of when different definitions would yield different competitive assessments. SCBC could endogenize some arbitrary choices that policymakers otherwise have to make.

Clustering to define markets is not new and appears in work such as Zwanziger *et al.* (1990), Lu (2017), and Ellickson *et al.* (2020). In the retail gasoline literature, clustering is emerging as a common tool to delineate markets (Carranza *et al.*, 2015; Lemus & Luco, 2021; Assad *et al.*, 2024; Martin, 2024). However, most of these applications cluster seller locations and do not use consumer locations directly.

To the best of our knowledge, our paper is the first to combine consumer location and seller identity when defining geographic markets with clustering. This approach lets SCBC capture market frictions that pure distance rules or administrative borders miss. These frictions are also emphasized in the new U.S. Department of Justice and Federal Trade Commission (2023) merger guidelines but not addressed in previous methods.

The new US DOJ merger guidelines describe several dimensions of geographic markets (U.S. Department of Justice and Federal Trade Commission, 2023, p.46):

Factors that may limit the geographic scope of the market includes transportation costs (relative to the price of the good), language, regulation, tariff and non-tariff trade barriers, custom and familiarity, reputation, and local service availability.

Our method can identify several of these dimensions in ways that a distance-

based algorithm cannot. For example, SCBC can capture natural boundaries, such as large roads, rivers, and mountains, when they affect consumer choices. In other settings, it may also detect patterns linked to segregation or discrimination.

SCBC relies on only two variables and does not require prices. This makes the approach useful in settings where demand-based methods are hard to implement, either due to data availability or due to the nature of the service. Examples include markets where consumers do not directly pay for the service, such as some platforms or privately provided public goods.

Like most observational methods, SCBC uses observed sales patterns. We do not observe how those patterns would change under counterfactual prices or quality changes. Instead, we use the realized market outcome, and this limitation also applies to related clustering approaches.

When validating our SCBC we use K-means (Macqueen, 1967), applied to consumer locations as a benchmark. K-means is one of the standard clustering algorithms and is used in related work (Yang, 2018; Ellickson *et al.*, 2020). SCBC connects to the probabilistic choice set definition, using information on the seller-consumer link in comparison to K-means. The requirement of probabilistic choice sets to be similar enough in theory is implemented and allows us to use data on observed sales patterns allowing seller to behave differently. Like all clustering methods, SCBC requires a choice of the number of clusters. This choice is often imposed externally, though some articles endogenize it from data (Carranza *et al.*, 2015). In our case, a standard Bayesian Information Criterion (BIC) is used to select the number of clusters. The BIC is a useful diagnostic, but should not be treated as a mechanical rule.

As an empirical exercise, we apply SCBC to the market for real-estate agent services for Stockholm apartments in 2017. This is a regulated market where

agent fees are small relative to property values, so price signals are less informative. Our data uses the apartment location as the consumer location and the agent used as the consumer-seller link. The setting is also policy-relevant. In 2015, the Swedish Competition Authority blocked a merger between two large real-estate agencies, and several strategies to improve competition have been discussed (Lind & Kopsch, 2014).

Using industry knowledge, we construct a validation set to evaluate SCBC. We find that SCBC classifies sales into markets significantly better than K-means. Areas separated by natural boundaries are often grouped together by K-means but separated by SCBC, without reducing accuracy in areas that are already well classified. The BIC points to five clusters, whereas the validation set suggests four, and when we assume five clusters we find a new pattern which almost perfectly matches the validation set, only splitting one of the validation clusters in two.

Market definition also affects measured concentration, captured by the Herfindahl-Hirschman index (HHI) in our data. K-means yields a lower HHI than both the validation set and SCBC, which suggests that a distance-only approach can understate concentration in policy work.

The article is structured as follows: Section 2 defines markets using choice sets, Section 3 describes SCBC, with technical details in Appendix B, Section 4 introduces the case study, Section 5 presents the results, Section 6 discusses the effect on market concentration, Section 7 gives a practitioners guide, and Section 8 concludes.

2 Choice Sets and Market Definitions

In this section, we define markets using choice sets for a homogeneous good. We define markets as sets of consumers. The dual definitions of market definition for

firms are presented in Appendix A. Given a homogeneous good, the definitions naturally extend to how geographic markets are delineated.

2.1 Primitives

Let $\mathcal{I}$ denote a finite set of consumers with $|\mathcal{I}| = N$, and let $\mathcal{J}$ denote a finite set of firms with $|\mathcal{J}| = J$. For consumer $i \in \mathcal{I}$, we define a mapping from each consumer to a set of firms that are feasible alternatives as the choice mapping $\mathcal{C}$.

$$\mathcal{C} : \mathcal{I} \rightarrow 2^{\mathcal{J}},$$

which implies the *choice set*

$$\mathcal{C}_i \subseteq \mathcal{J},$$

which represents the set of firms that are feasible alternatives for consumer i .

From the firm perspective, each firm is associated with a set of all the consumers whose choice mapping maps into that firm $j \in \mathcal{J}$. This can be thought of as the inverse image of the choice mapping and induces the *exposure set*

$$\mathcal{C}_j^{-1} \subseteq \mathcal{I},$$

which represents the set of consumers who have firm j in their choice set.¹

2.2 The definition of markets with identical choice sets

In a strict setting, we allow no differences among consumers in the same market. We define the markets as follows.

Definition 1 (Strict Consumer Market) *A Strict Consumer Market is a*

¹This is the pre-image of the choice mapping $\mathcal{C}$.

subset of consumers $\mathcal{I}_g \subseteq \mathcal{I}$ such that

$$\mathcal{C}_i = \mathcal{C}_{i'} \quad \forall i, i' \in \mathcal{I}_g,$$

and no strict superset of $\mathcal{I}_g$ satisfies this property.

This definition states that a market is composed of all consumers who have the same choice set. Consumers in the exposure set of one firm might, however, be in different markets. Any deviation from the choice set is treated as a new market, even if it reflects only the exclusion of one firm.

2.3 The definition of markets with connectedness

A broader definition instead considers connectedness. No consumer considers firms in more than one market, and no firm serves consumers in different markets. We define *Connected Consumer Markets* as follows.

Definition 2 (Connected Consumer Market) A Connected Consumer Market is a subset of consumers $\mathcal{I}_g \subseteq \mathcal{I}$ such that for any partition $A \subset \mathcal{I}_g$, with $A \neq \emptyset$ and $A \neq \mathcal{I}_g$,

$$\bigcup_{i \in A} \mathcal{C}_i \cap \bigcup_{i \in \mathcal{I}_g \setminus A} \mathcal{C}_i \neq \emptyset,$$

and no strict superset of $\mathcal{I}_g$ satisfies this property.

This definition says that a market is all consumers who are linked by the choice mapping. If one consumer can be linked to another through repeated application of the choice and exposure relation, they belong in the same market. With this definition, the dual definition presented in Appendix A will yield the same markets.

2.4 Defining markets probabilistically

To allow for some overlap between both the consumers' choice sets and the firms' exposure set, we take a probabilistic approach. We replace the choice mapping $\mathcal{C}$ with the probabilistic choice mapping $\mathcal{T}$.²

$$\mathcal{T} : \mathcal{I} \rightarrow [0, 1]^{\mathcal{J}}.$$

We represent the probabilistic choice mapping with a matrix $\mathbf{\Pi} \in [0, 1]^{N \times J}$ with elements $\pi_{ij} = \mathbb{P}(j \in \mathcal{C}_i)$. Each element is the probability that firm j is in the choice set of consumer i . We can extract the row elements to get the equivalent of the probabilistic choice set, $\mathcal{T}(x_i)$, for the consumer i .

$$\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{iJ}) \in [0, 1]^J.$$

If we extract the columns, we get the equivalent of the exposure set:

$$\boldsymbol{\pi}_j = (\pi_{1j}, \dots, \pi_{Nj}) \in [0, 1]^N.$$

The deterministic exact choice mapping is a special case when $\pi_{ij} \in \{0, 1\}$ for all i and j .

We can now define the *Strict probabilistic Consumer Market*:

Definition 3 (Strict probabilistic Consumer Market) *A Strict probabilistic Consumer Market is a subset of consumers $\mathcal{I}_g \subseteq \mathcal{I}$ such that*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_{i'} \quad \forall i, i' \in \mathcal{I}_g,$$

and no strict superset of $\mathcal{I}_g$ satisfies this property.

² $\mathcal{C}$ is a special case of $\mathcal{T}$ with $\mathcal{T} : \mathcal{I} \rightarrow \{0, 1\}^{\mathcal{J}}$

This allows some firms to serve more than one market at the same time, as it allows consumers to consider firms in different markets, as long as the probability of consumers in one market considering every unique firm is the same.

As π_i is identical within a market g , $\pi_i = \pi_g$, $\mathbf{\Pi} \in [0, 1]^{N \times J}$ can then be reduced to $\mathbf{\Pi} \in [0, 1]^{G \times J}$. This is the equivalent of $\mathcal{T}(x_i) = \mathcal{T}(g) \quad \forall i \in \mathcal{I}_g$.

2.5 Geographic market definitions

In geographic markets for a homogeneous good, the choice mapping is a function of the location. If we write $x_i \in R^2$ as the location of consumer i , and $x_j \in R^2$ as the location of firm j and a distance metric $d(x_i, x_j)$. If we assume that consumer i considers all firms within distance δ , we can define the consumer market as

$$j \in \mathcal{C}(x_i) \text{ if } d(x_i, x_j) \leq \delta$$

The distance-based definition mimics the methods used in the literature when a consumer is assumed to be in the market of a firm if they are within a distance δ . When defining two firms as competitors if they are within a distance δ the same logic is followed.

For the probabilistic choice mapping, we can write the probability that firm j is in consumer i 's choice set as a function of the distance between x_i and x_j , $\mathcal{T}(x_i)(d(x_i, x_j)) \quad \forall j \in \mathcal{J}$.

Such a definition is useful when the location of the firm is clear and relevant for the choice of the consumer. In some cases, the location of the firm might be less relevant, for example when the good is delivered to the consumer, or when the location of the firm is not known. In those cases, we can write $\mathcal{C}(x_i)$ for the consumer-based market, where the choice set depends on consumer location whereas firm location itself is irrelevant. This is a generalization of the distance-

based choice mapping. The same applies to the probabilistic choice mapping, $\mathcal{T}(x_i)$.

The empirical challenge is then to recover the choice mapping, or the probabilistic choice mapping, from data. Because the exact choice mapping is never observed, it is easier to work with the probabilistic choice mapping for tractability. We will discuss this in the next section.

2.6 Empirical probabilistic choice mapping

If we assume a homogeneous good and that the probabilistic choice mapping is a function of the location $\mathcal{T}(x_i)$, we can assume that $\mathcal{T}(x_i) \approx \mathcal{T}(x_{i'})$ if $d(x_i, x_{i'})$ is small. We will intuitively expect that consumers that are close to each other will have similar choice sets. The intuition is that we proxy consumer i 's choice set by the choice sets of nearby consumers. If we assume that the choice mapping is a continuous function of the location, any consumer i will have slightly different probabilistic choice mapping than nearby consumers. Applying definition 3 strictly would then yield one market per consumer or all consumers in one market. We therefore need to update Definition 3

Definition 4 (Probabilistic Consumer Market) *A Probabilistic Consumer Market is a subset of consumers $\mathcal{I}_g \subseteq \mathcal{I}$ such that*

$$\pi_i \sim \pi_{i'} \quad \forall i, i' \in \mathcal{I}_g,$$

subject to the condition C to decide which is $\sim$ enough. Further, no strict superset of $\mathcal{I}_g$ satisfies this property.

The condition C can be set in different ways.

Remembering that in the definition 3 $\mathcal{T}(x_i) = \mathcal{T}(g) \quad \forall i \in \mathcal{I}_g$, where $\mathcal{T}(g)$ is the probabilistic choice mapping of market g . If we instead treat $\mathcal{T}(g)$ as

the representative probabilistic choice mapping of the consumers in market g , we can then define a distance metric between the probabilistic choice mapping. For example, we can use the L_2 distance to define the condition C , and say that consumers with a distance below a certain threshold from $\mathcal{T}(g)$ belong to the market g . In the matrix notation this becomes $i \in \mathcal{I}_g$ if $\|\pi_i - \pi_g\| \leq \eta$ where π_g is the representative probabilistic choice mapping of market g and η is the threshold of similarity. Such a definition allows for an individual to be in multiple markets, which might not always be a desirable feature.

Another way is to impose a number of markets and see how we can split R^2 in a way that minimizes an objective function as the condition C .³ To do so we will use a clustering method.

3 Spatial and Categorical Bayesian Clustering (SCBC)

In Section 2, we define a probabilistic, consumer-based market concept using latent choice environments. The empirical task is to recover these latent geographic markets from observed data. We will do so by using clustering. First we will cluster consumers based only on location, as in much of the existing literature, and then extend the framework to include seller identities through SCBC.

Traditional delineation of markets often relies on administrative borders or ad hoc distance rules. More recent work uses statistical clustering methods.

³An illustrative example of an objective function would be the within-market variance of the probabilistic choice mapping.

3.1 Clustering

Clustering partitions observations into groups. A useful distinction is between deterministic (hard) and model-based (soft) methods (Macqueen, 1967; Hastie *et al.*, 2001). Hard methods assign each observation to one cluster; soft methods assign membership probabilities across clusters. Gaussian mixture models (GMMs) are a standard model-based soft approach.

In market delineation, commonly used methods include K -means (Ellickson *et al.*, 2020; Yang, 2018) and hierarchical clustering (Assad *et al.*, 2024; Carranza *et al.*, 2015; Lemus & Luco, 2021).

The K -means algorithm is deterministic: it minimizes within-cluster variation under a distance metric, typically Euclidean distance (Athey & Imbens, 2019), by assigning each observation x_i to the nearest centroid.

Model-based clustering instead treats each cluster g as a probability distribution and assigns observations based on posterior cluster probabilities (Fraley & Raftery, 2002; Lavine & West, 1992; McLachlan *et al.*, 2019). Mixture models, especially GMMs, are the most common specification and allow out-of-sample assignment of new observations.

3.2 Gaussian Mixture Model

The model-based analogue to K -means is the Gaussian mixture model (GMM). A GMM represents each cluster as a Gaussian distribution and produces soft assignments through posterior responsibilities (Fraley & Raftery, 2002; Lavine & West, 1992; McLachlan *et al.*, 2019).

A Gaussian mixture model (GMM) is characterized by a mixture density

$$f(x_i | \Phi) = \sum_{g=1}^G \omega_g \mathcal{N}(x_i | \boldsymbol{\mu}_g, \Sigma_g),$$

where $\Phi = \{(\omega_g, \boldsymbol{\mu}_g, \Sigma_g)\}_{g=1}^G$ denote the parameter set that describes the model. The G distributions have the center $\boldsymbol{\mu}_g \in \mathbb{R}^2$ and covariance matrix $\Sigma_g \in \text{PD}_2$. $\omega_g \geq 0$ are the mixing proportions and $\sum_{g=1}^G \omega_g = 1$. We also define the posterior responsibilities as $p(z_i = g \mid x_i, \Phi)$.

A GMM is fitted to data by estimating Φ , this is done with an EM algorithm. Once we have estimated Φ , we classify any point $x \in \mathbb{R}^2$ using the maximum a posteriori (MAP) rule:

$$g' = \arg \max_{g \in \{1, \dots, G\}} p(z_i = g \mid x_i, \Phi).$$

GMM can categorize any continuous data into clusters. We will use it to categorize consumer locations into clusters. The clusters will be used as the latent geographical markets.

3.3 Fitting a GMM with categorical data: SCBC

When a standard GMM is used for clustering, estimation (typically via EM) yields posterior responsibilities $p(z_i = g \mid x_i, \Phi)$ for each observation i and cluster g :

$$x_i \mid z_i = g \sim \mathcal{N}(\boldsymbol{\mu}_g, \Sigma_g).$$

This captures purely spatial clustering and estimates the mixture parameters Φ . Under this baseline, market approximation depends only on location as in Section 2.5.

SCBC extends this setup with a Dirichlet-GMM (DGMM), using the consumer-seller link to incorporate common choice environments in market assignment. Because individual choice sets are unobserved, we use realized transaction patterns as proxies.

Let y_i denote the seller identity for consumer i . Conditional on y_i , the

responsibility for consumer i follows

$$z_i \mid y_i \sim \text{Categorical}(\theta_y).$$

The seller distribution over markets vector is given a Dirichlet prior:

$$\theta_y \sim \text{Dir}(\alpha),$$

with symmetric hyperparameter α , so the data determine the degree of seller concentration. The vector θ_y is a normalized observable proxy for the exposure set in the representative probabilistic choice mapping Π in Section 2.

The posterior assignment rule maximizes

$$\Pr(z_i = g \mid x_i, y_i, \Phi) \propto \phi(x_i \mid \mu_g, \Sigma_g) \cdot \theta_{g, y_i},$$

instead of

$$\Pr(z_i = g \mid x_i, \Phi) \propto \phi(x_i \mid \mu_g, \Sigma_g).$$

Including θ_y adds seller-based information to the assignment, so clusters are shaped jointly by geography and seller distribution.

3.4 Implementing SCBC with a DGMM

We observe N consumers and J sellers. The data are pairs $\{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^2$ is the consumer location and $y_i \in \{1, \dots, J\}$ is the realized seller. Each observation has a latent assignment $z_i \in \{1, \dots, G\}$.

Each seller has a market concentration vector $\theta_j = (\theta_{1j}, \dots, \theta_{Gj})$ where θ_{gj} is the probability that seller j appears in market g , $\theta_{gj} \geq 0$, and $\sum_{g=1}^G \theta_{gj} = 1$. By incorporating seller patterns into clustering, SCBC combines demand-side outcomes with spatial structure. This is the main difference from earlier

location-only clustering approaches to market definition.

The theoretical motivation for sellers’ concentrating on a few markets in defining markets can be found both in (Elzinga & Hogarty, 1973) and the merger guidelines from the U.S. Department of Justice and Federal Trade Commission (2023).

3.5 Implementation

As in Bayesian mixture models, cluster-specific densities define the classification likelihood. With model parameters Φ and data $\mathbb{X}$, the likelihood is

$$\mathcal{L}(\Phi, \mathbb{X}) = \prod_{i=1}^N \sum_{g=1}^G (p(X|\Phi)p(z_i|\Phi)) \quad (1)$$

Appendix B presents the full model specification, priors, and likelihood.

The target is the posterior distribution over latent clusters and parameters. Because this posterior is not analytically tractable, we use Markov chain Monte Carlo, specifically a Gibbs sampler (Metropolis *et al.* , 1953; Hastings, 1970; Geman & Geman, 1984; Bernardo & Smith, 1994; Lavine & West, 1992). The chain is constructed to have $p(z, \mu, \Sigma, \Theta \mid x, y)$ as its invariant distribution.⁴

Once the Gibbs sampler has converged, we have an estimated posterior to sample from for the clusters g (i.e. their locations and shapes) and can assign any location $x \in \mathbb{R}^2$ to a cluster using an updated MAP rule:

$$g' = \arg \max_{g \in \{1, \dots, G\}} p(z_i = g \mid x_i, y_i, \Phi).$$

3.6 The number of clusters

The drawback of any clustering algorithm is that the number of clusters G is set exogenously. Different approaches to endogenously select the number of clusters

⁴See Appendix B for the full derivation.

have been made. The most common one is the Bayesian information criterion (BIC) (Bernardo & Smith, 1994; Raftery, 1995), based on the Bayes factor. It is derived from the likelihood of the distribution and the number of parameters. The BIC is defined as:

$$\text{BIC} = (H + J) * G * \ln(N) - 2 * \ln(\mathcal{L}(z|x, y)), \quad (2)$$

where H is the number of parameters that pin down each cluster, J is the number of sellers, G is the number of clusters and N the number of consumers. The rule is to select G such that the model estimated minimizes BIC. We iterate the algorithm over G , and benchmark models with a low BIC score as appropriate.

3.7 Validation set

In order to assess the validity of the model, we need a validation set. The validation set will correspond to the correct market delineation. But as we do not know the correct market delineation, we use industry knowledge to approximate it. In Section 4.1, we will present how we create the validation set for our empirical example. In a setting where the market is well known, like our empirical example, we can establish how well the model performs in market delineation. And once the performance is established, the method can be used in settings where there is no commonly agreed market delineation and therefore no validation set.

We use the Dice coefficient to evaluate SCBC. It spans between 0 and 1, where a higher value indicates a better classification. The Dice coefficient is calculated as the overlap between the validation set and the outcome of the clustering algorithm. It is a simple metric but gives us a sense of how well the algorithm performs.

4 An empirical example, the Stockholm real estate agents market

4.1 Setting

In our empirical example we study the market for real-estate agent services in downtown Stockholm. We use apartment transaction data to recover local choice environments by applying SCBC.

In Sweden, real-estate agents represent the seller of the apartment, and commissions are low in an international comparison.⁵ Real-estate agents also commonly operate through franchise offices with designated territories and non-compete restrictions vis-à-vis other offices in the same chain. Real estate agents in Sweden most often work for large franchise chains. The franchisee has its own office and coverage area. They have a non-compete agreement with other offices of the same franchise. They are therefore often restricted to selling in a designated area. These institutional features make seller identity informative about the latent set of agents that are feasible alternatives in a given area.

Stockholm’s geography is characterized by being built on islands. It consists of four well-known districts on three islands: Kungsholmen, Södermalm, Vasastan, and Östermalm. Two of the districts are their own islands, whereas the other two share parts of the mainland, but still surrounded by mostly water. The four districts are historically administrative regions and strong identity markers in the city, although the two administrative districts that are on the mainland were recently merged. Further, when listing a real estate property, the name of the district is highly salient, and buyers usually have clear preferences over districts. Taken together with industry knowledge, these features make us comfortable using the listing areas as a credible validation set.

⁵3.5% without tax, compared to 5% and 7.7% in Finland and the US respectively according to Lind & Kopsch (2014).

Because the validation set in the Stockholm case provides a credible market delineation, this setting is well suited to testing a data-driven method for delineating markets. The aim of the empirical exercise is to evaluate whether the use of seller identity and location together in SCBC captures shared choice environments better than a distance-based clustering algorithm. If the predictive power is better in this empirical setting it strengthens the argument that SCBC can be used by practitioners to remove ad hoc choices when delineating markets.

4.2 Data

The sample used for the analysis is sales of apartments in the inner Stockholm area from 2017.⁶ We use all online listed sales in the four districts in downtown Stockholm and restrict the sample to agents who sell at least 10 apartments during the period, which leaves us with a sample of 3145 observations.

The relevant variables for the estimation are intentionally sparse. For each apartment we observe the geographical position of apartment i , $(x_i \in \mathbb{R}^2)$ and the real-estate agent who sold it, (y_i) . These are the only variables we use to estimate our model and recover the latent market assignment z_i . We use the official listing area to validate our method, and it is not used in estimation. To evaluate the gain from including seller identity in the estimation, we compare how well SCBC overlaps with the validation set relative to a distance-only method. We here use standard K -means clustering, which uses location alone.

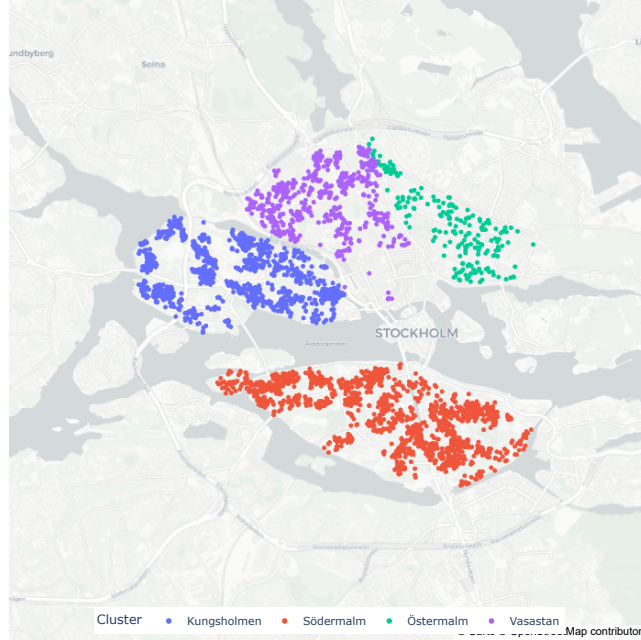
5 Results

The novelty of SCBC is the addition of seller identity to a purely location-based method of market definition. The empirical question is whether this addition

⁶We are grateful to Petter Berg for supplying the data.

helps recover latent geographic markets, as it should better reflect shared choice environments. We begin by plotting the validation set in Figure 1, where each color corresponds to one of the four listing areas. These listing areas are not treated as the truth by assumption, but as an institutionally grounded validation set for whether the estimated boundaries predict a proxy for the correct latent geographic market.

Figure 1: Downtown Stockholm plotted with the validation set



Notes: The dataset of the empirical example covering downtown Stockholm. The different colors correspond to the different values of the validation set. The validation set is the four listing areas Kungsholmen, Södermalm, Vasastan, and Östermalm. Note that Kungsholmen and Södermalm are two islands and Vasastan, and Östermalm are on the mainland, but delineated by a strait line.

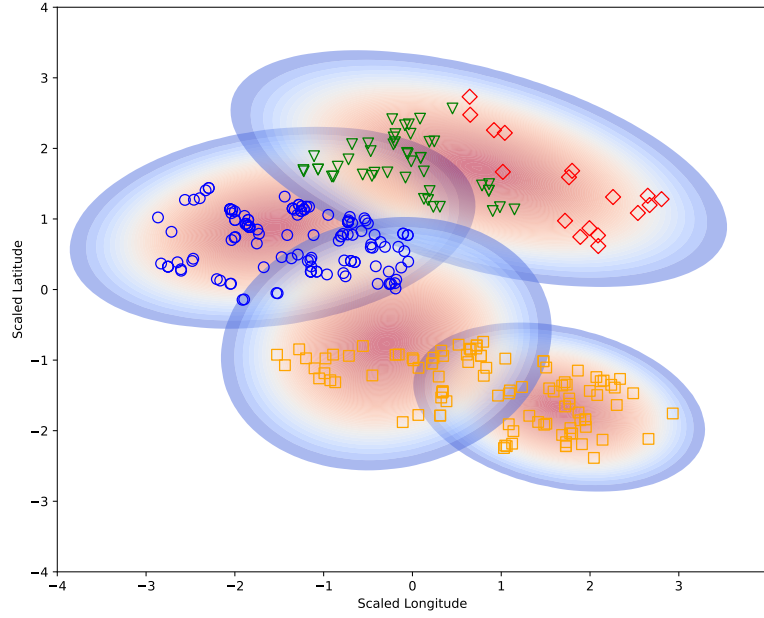
We first estimate the baseline K -means algorithm to fit four clusters to the sample data. SCBC needs to have better predictive power than the benchmark for it to be a useful tool. Although the K -means method is not a Gaussian mixture, we can extract an implied Gaussian mixture from the cluster assignments.

This will be the prior for SCBC when we implement it with a Dirichlet Gaussian Mixture Model (DGMM). We then show how these distributions change as we apply the DGMM.

5.1 Mechanisms of the method

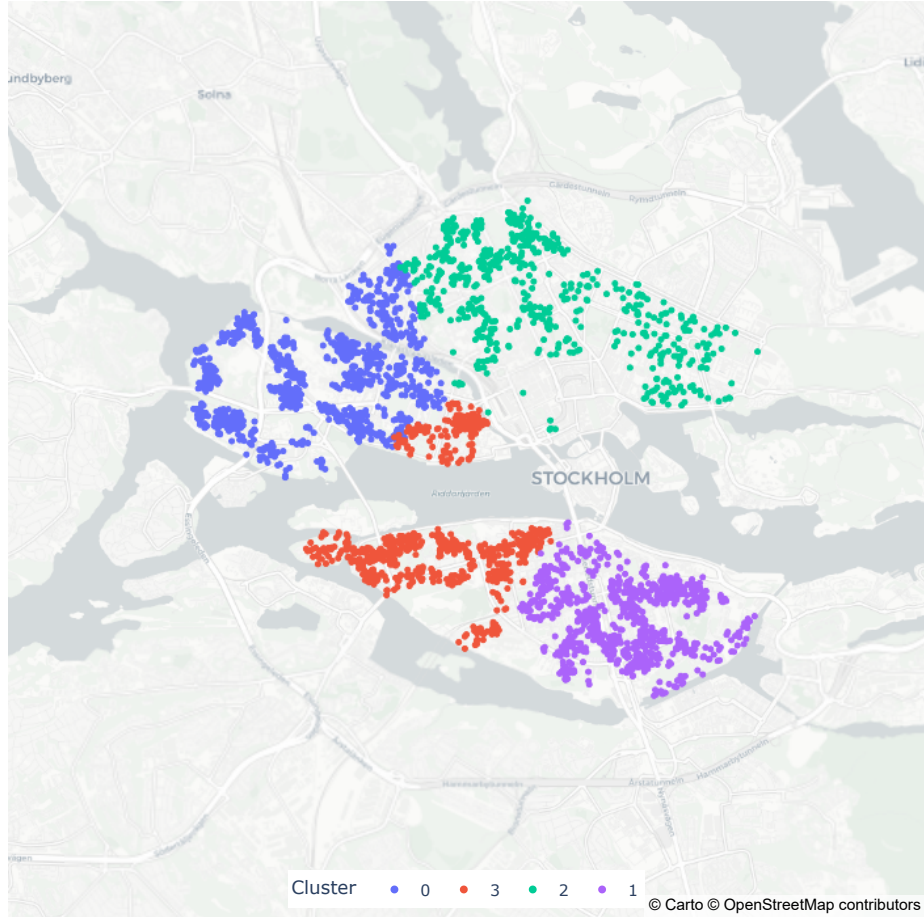
We extract implied Gaussian distributions over $\mathbb{R}^2$ from the K -means clustering assignment. They are defined by the center of each distribution μ_g and the covariance matrix Σ_g . The implied cluster distributions are plotted in Figure 2, where each of the shaded areas corresponds to a Gaussian probability distribution. A sample of the observations is plotted on top of the distribution. The different shapes of the observations indicate which validation cluster they belong to. When we apply the classification using the DGMM, each observation will have a cluster which is “responsible” for it. That is the cluster that has the highest probability for the observation’s location. We say that the observation is assigned to that cluster. The substantial overlap of the distributions indicates the challenge in delineating the markets. The similarity in size and shape of the probability distributions seems to increase the overlap. Visually, the distributions do not clearly separate the validation set. For example the blue circles seem to be split. The left distribution seems to be responsible for the majority, but the middle distribution is responsible for a substantial minority.

Figure 2: Distributions of the Gaussian distribution of the K-means clusters



Notes: The data points are observations with the different shapes and colors indicating the validation set, i.e. the different listing areas. The distributions are the Gaussian distributions of each of the four clusters that correspond to the different markets. The distributions in this figure correspond to the baseline K-means algorithm.

Figure 3: Classification of the data points using the baseline K-means algorithm



Notes: The dataset of the empirical example covering downtown Stockholm. The different colors correspond to the different values of classification using the baseline K-means method. Note how the clusters do not overlap with the validation set and do not follow the natural borders.

Figure 3 shows the classifications made by the benchmark K -means method. Here the western part of Stockholm, Kungsholmen is split into two markets, such that the eastern part of the island is instead linked with the southern island of Södermalm. The limitations of the distance-only method become evident as it is at odds with the validation set in Figure 1. When using purely distance, the institutional and geographic features that shape local choice sets are not

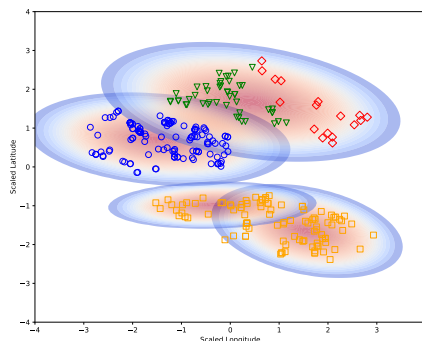
captured.

We next turn to SCBC, where the method includes not only the location data but also the seller link. If a real-estate agent sells in a location, this is informative about the common choice environments around that location. They have a higher probability in the probabilistic choice set of not only the consumer but also their neighbors. When SCBC iterates the DGMM it updates the covariance matrix Σ_g and the centers μ_g of the cluster.

In Figure 4 the same data is plotted as in 2, but now the distributions are the SCBC distributions. The new distributions have a larger variation in shape and are more consistent with natural barriers and local specialization visible in the validation set.

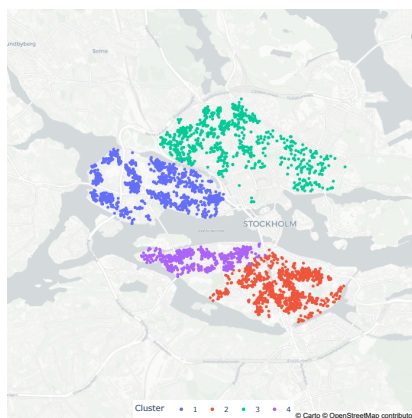
The classification is shown in Figure 5. Relative to K -means, the western island of Kungsholmen is now assigned to one market. SCBC has tightened and reshaped the distributions by incorporating the common choice environment as proxied by the seller link. Consumers close to a border but on the same side face a similar choice environment, whereas consumers close to a border but on different sides do not. As the choice environments are not observable, the seller link helps proxy the latent choice environments as agents likely operate only on one side of the border.

Figure 4: Distributions for the SCBC



Notes: The data points are observations with the different shapes and colors indicating the validation set, i.e. the different listing areas. The distributions are the Gaussian distributions of each of the four clusters that correspond to the different markets. The distributions in this figure correspond to the SCBC method.

Figure 5: Classification using SCBC

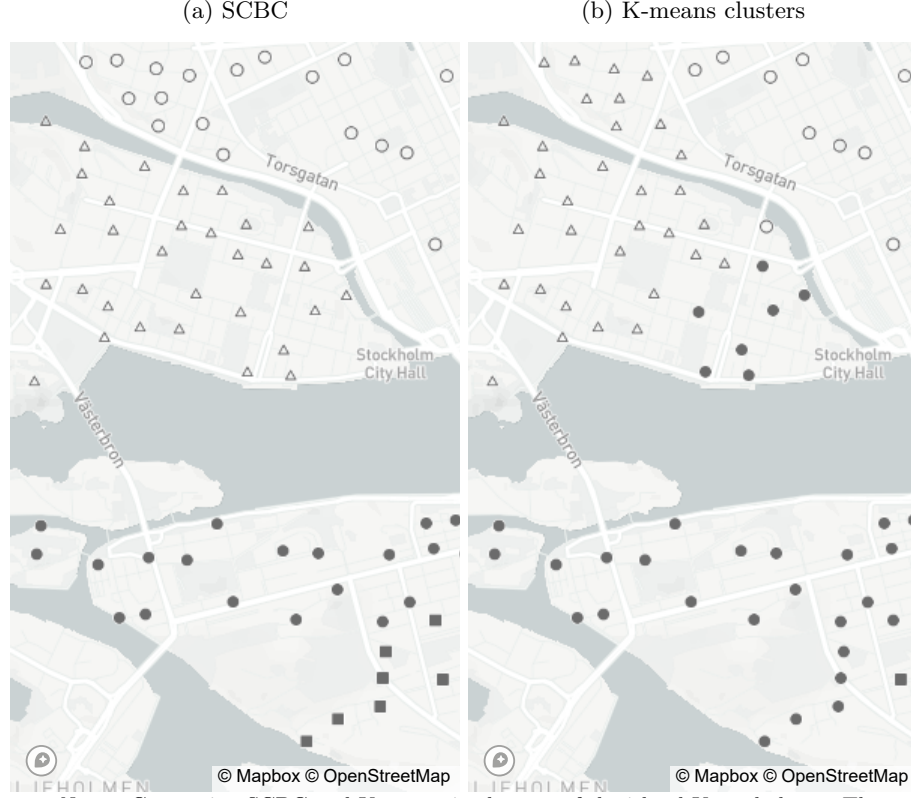


Notes: The dataset of the empirical example covering downtown Stockholm. The different colors correspond to the different values of classification using SCBC. Note how the clusters follow the natural borders, but do not match the validation set entirely.

Figure 6 zooms in on one of the natural boundaries. There, SCBC helps explain the disagreement regions. There, SCBC separates observations along the water, whereas K -means assigns nearby locations on opposite sides of the

barrier.

Figure 6: The clustering by SCBC compared to K-means on a map



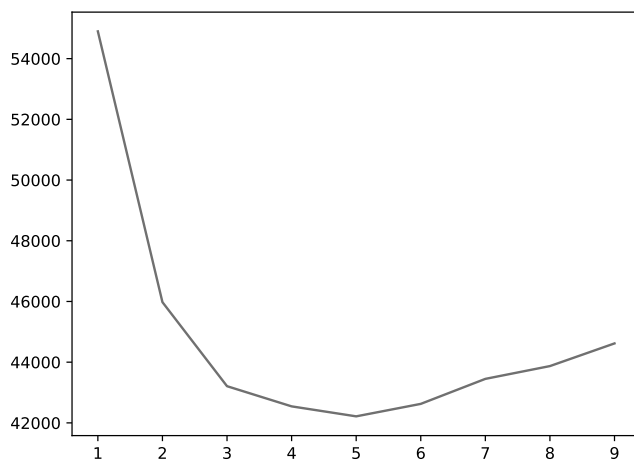
Notes: Comparing SCBC and K-means in the east of the island Kungsholmen. The different shapes indicate different clusters.

To check the validity, we compute the Dice scores of SCBC and K -means with regard to the listing areas which we treat as the validation set. Both algorithms approximate the underlying market structure to some degree. The Dice score is 0.78 for SCBC and 0.67 for K -means, implying that incorporating seller identity brings the estimate closer to institutional knowledge. The use of four clusters is the most consistent way to validate the methods. However, this is a conservative approach, and varying the number of clusters can materially improve SCBC predictions.

5.2 The number of clusters

In clustering methods, the number of clusters often needs to be set exogenously. Testing different numbers of clusters and calculating the BIC allows for evaluation of the fit, where a lower BIC indicates a better fit. Figure 7 reports the BIC scores for different numbers of clusters. Five clusters yield the lowest BIC, not four as expected based on institutional knowledge.

Figure 7: The BIC value for different numbers of clusters

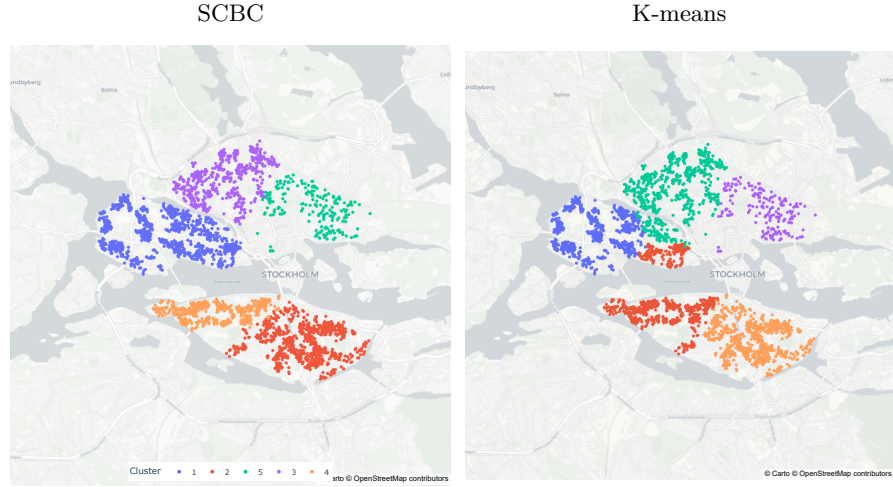


Notes: The BIC test calculates the BIC for different numbers of clusters and compares which number of clusters have the lowest value and is therefore the theoretical correct number of clusters. In this example, the BIC test indicates that five is the correct number of clusters. Note however that the difference compared to the value for four and six clusters is small.

Assuming five clusters uncovers a new pattern, where the northern clusters are divided into two. SCBC divides the observations along a line similar to that in the validation set while not altering the adaptation to the natural borders. *K*-means also splits the clusters, but neither along the natural boundaries nor along the border of the validation set. The five-cluster exercise results in a split broadly consistent with the validation set, with the exception that the southern

cluster is divided in two. In general, this exercise suggests that there might be finer market structures of importance that are revealed using the choice-set method, but not captured in the institutional knowledge. SCBC may therefore be a tool to identify which areas to investigate further.

Figure 8: SCBC compared to K-means for five clusters



Notes: The clustering by SCBC compared to K-means, assuming five clusters instead of four. The symbols indicate different clusters. Note that for SCBC we have the same three markets for the two islands Kungsholmen and Södermalm. The cluster that used to be Vasastan and Östermalm is now split into two, close to the split in the validation set. The K-means method once again splits the data into clusters that do not follow the validation set.

6 Measure of market concentration

The choice of market delineation has implications for the implied intensity of competition. An incorrect assignment of observations from two distinct markets into the same market will increase the perceived competition even if the underlying competitive environment stays the same. Mechanically, SCBC fits clusters so that the sales for each real-estate agent will be concentrated in as few markets as possible. Therefore, they appear more concentrated than under K-means and closer to the validation set.

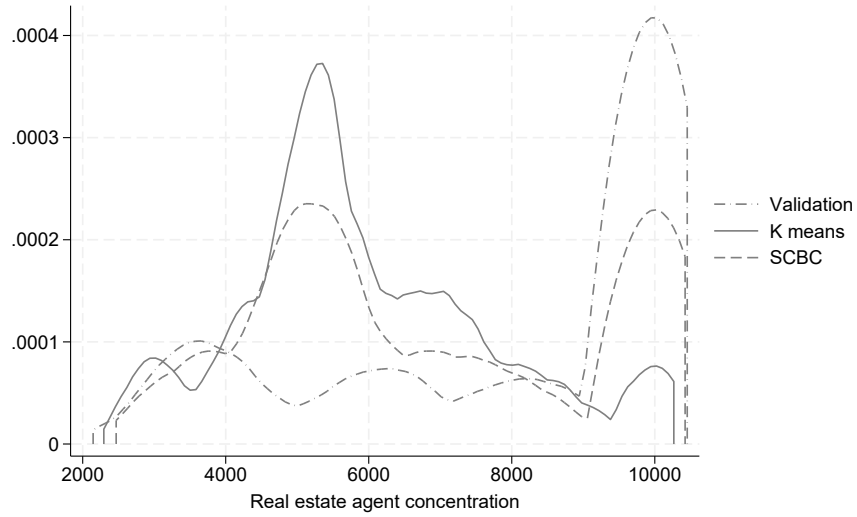
For each real-estate agent l , we compute a concentration measure, Con_l . It is defined as the sum of squared sales across markets divided by squared total sales:

$$Con_l = \sum_k^K \frac{Sales_k^2}{Total\ Sales^2}.$$

Figure 9 presents the distribution of this concentration measure for our sample. The three distributions correspond to the three methods of market delineation.

SCBC identifies more real-estate agents with high concentration than K -means and the SCBC distribution is closer to that of the validation set. When market boundaries align more closely with latent choice environments, agents appear active in fewer distinct markets. By contrast, the distance-only partition cuts across local boundaries and implies that agents work in multiple markets.

Figure 9: The concentration of each real estate agent, ranging from 0 to 10 000



Notes: In this graph we see a kdensity plot of each agent concentration Con_l for the three different specifications. The concentration is measured as the HHI index of each agent over the four markets. A concentration of 10 000 is equivalent to all sales in one market. We use a triangular kernel and 100 estimation points. Note that the maximum value is 10 000, but that some data points are above that due to the kernel density estimation.

To investigate whether these differences have economic implications, we also study the Herfindahl-Hirschman Index (HHI) under the different methods for market definition. Using the validation set as the market definition yields a weighted average HHI of 942. The corresponding value is 850 when using SCBC and 827 when using K -means. SCBC results in a HHI measure that is closer to the market concentration implied by using the validation set than the distance-only definition, K -means. The differences are modest, and by conventional HHI thresholds the Stockholm real-estate-agent market is not highly concentrated.

Table 1: The HHI for the four different markets

Market	Validation Set	K-means	SCBC
1	793	739	702
2	802	747	790
3	963	917	926
4	1938	920	1064

Notes: The table displays the HHI between 0 and 10 000 for the four different markets in 2017, using the three different methods.

7 Practitioners' guide

In this article, we formulate a choice-set approach to market definition and implement it using SCBC in the Stockholm setting to compare estimated markets with a validation set. However, the broader aim of the article is to develop a framework and method, SCBC, that can delineate markets in the absence of a validation set by using information on consumers' choice environments. For a practitioner, the objective is not to replicate a known partition, but to better understand the market boundaries that are economically relevant when assessing competition.

To apply SCBC in practice, three main concerns should be addressed: the number of clusters, confidence in the results, and the computational power required. Relying on priors based on institutional knowledge can help guide both

the choice of the number of clusters and the reliability of the resulting delineation.

When choosing the number of clusters, a sensible starting point is the best guess based on prior knowledge. Applying the BIC while varying the number of clusters around that initial guess can help refine it by identifying splits or aggregations that better fit consumers' feasible choices. Applying the BIC globally and choosing the lowest possible value would be the theoretically correct approach, but it risks aggregating all markets into one or overfitting such that definition is one market per observation. It would also be computationally intensive.

SCBC requires a prior. In the standard method, *K*-means provides that prior. When available, a prior based on industry knowledge may be a better alternative. As with the number of clusters, the market delineation produced by SCBC should be compared with industry knowledge, and deviations that appear nonsensical should be taken as a warning. In such cases, the results should be interpreted with care, just as with the BIC diagnostic.

Consider a hypothetical example in which industry knowledge suggests delineating an area into three geographic markets, but there is no consensus on where the borders lie. Applying SCBC with the number of clusters set to three would then yield a natural baseline. If the three-cluster specification indicates that two of the expected markets should be merged whereas the third should be split into two clusters, further investigation would be warranted.

The SCBC result could have different explanations. It may be that, once choice environments are taken into account, the industry view is misleading. Alternatively, the market may not consist of three distinct markets, but rather two or four. Comparing the three-cluster specification with two- and four-cluster specifications would help the analyst distinguish between these interpretations.

A two-cluster specification would indicate a coarser market definition than the industry view, whereas a four-cluster specification would indicate a finer one. If they align on the relevant areas, the remaining question is which level of aggregation is most appropriate. Here, the BIC diagnostic could be useful in indicating the most plausible number of clusters.

Finally, implementation costs should also be taken into account. SCBC is implemented with a Dirichlet Gaussian mixture model, and the computational estimation of that model is carried out with a Gibbs sampler. This method is implemented in Python, and the computational power required can be substantial even for small datasets. When implementing SCBC in a new setting, these computational demands should be considered, and it may be necessary to use a subsample of the available data.

8 Conclusion

This article makes two contributions. First, it develops a choice-set framework for geographic market definition in which markets are understood as groups of consumers facing similar feasible alternatives. That framing links market delineation directly to substitution opportunities rather than to administrative borders or purely ad hoc distance rules. Second, it shows how this idea can be implemented empirically with a tractable clustering procedure that combines location with seller identity.

The procedure developed, Spatial and Categorical Bayesian Clustering (SCBC), has modest data requirements and can therefore be applied in many settings. In the Stockholm real-estate-agent market, this empirical approach recovers boundaries that are closer than those of a distance-only benchmark to the validation set derived from institutional knowledge. Seller identity contains information about common choice environments, and that information helps explain the

improvement.

Having a method to suggest geographic market definitions in addition to existing methods would be useful, both for competition authorities and private companies. SCBC is adaptable to different settings and requires relatively little data compared with other methods that are not purely distance based. After further testing, it could complement existing approaches. At its current stage, SCBC should be tested in other settings in which customer location is central to geographic market delineation, such as online shopping, dentists, doctors, and other local services with limited data availability. SCBC may also help identify patterns of discrimination and segregation.

SCBC should be interpreted as an empirical approximation of consumers' choice environments. Its usefulness depends on whether realized seller-identity patterns, in combination with location, are informative about latent choice sets. With these caveats, both the choice-set framework and the method can be valuable in settings where prices are unavailable, institutional boundaries are disputed, and geographic market definition remains necessary for applied competition analysis.

References

- Assad, Stephanie, Clark, Robert, Ershov, Daniel, & Xu, Lei. 2024. Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. *Journal of Political Economy*, **132**(3), 723–771.
- Athey, Susan, & Imbens, Guido W. 2019. Machine Learning Methods That Economists Should Know about. *Annual Review of Economics*, **11**(March), 685–725.

- Bernardo, J.M., & Smith, A.F.M. 1994. *Bayesian Theory*. Chichester: John Wiley and Sons.
- Carranza, Juan Esteban, Clark, Robert, & Houde, Jean François. 2015. Price controls and market structure: Evidence from gasoline retail markets. *Journal of Industrial Economics*, **63**(1), 152–198.
- Competition and Markets Authority. 2015 (September). *Poundland and 99p: Final Report on the Anticipated Acquisition by Poundland Group plc of 99p Stores Limited*. Issued by the Competition and Markets Authority, UK.
- Ellickson, Paul B., Grieco, Paul L.E., & Khvastunov, Oleksii. 2020. Measuring competition in spatial retail. *RAND Journal of Economics*, **51**(1), 189–232.
- Elzinga, Kenneth G., & Hogarty, Thomas F. 1973. The Problem of Geographic Market Delineation in Antimerger Suits. *The Antitrust Bulletin*, **18**(1), 45–81.
- Elzinga, Kenneth G., & Howell, Vandy M. 2018. Geographic Market Definition in the Merger Guidelines: A Retrospective Analysis. *Review of Industrial Organization*, **53**(3), 453–475.
- Fraley, Chris, & Raftery, Adrian E. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.
- Geman, Stuart, & Geman, Donald. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**(6), 721–741.
- Genakos, Christos, & Pagliero, Mario. 2022. Competition and Pass-Through: Evidence from Isolated Markets. *American Economic Journal: Applied Economics*, **14**(4), 35–57.

- Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.
- Hastings, Author W K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**(1), 97–109.
- Lavine, Michael, & West, Mike. 1992. A Bayesian method for classification and discrimination*. *The Canadian Journal of Statistics*, **20**(4).
- Lemus, Jorge, & Luco, Fernando. 2021. Price Leadership and Uncertainty About Future Costs. *Journal of Industrial Economics*, **69**(2), 305–337.
- Lind, Hans, & Kopsch, Fredrik. 2014. *Konkurrensen på fastighetsmäklarmarknaden*. Tech. rept. Konkurrensverket, Stockholm.
- Lu, Anna W. 2017. *Three Essays on Empirical Industrial Organization in Grocery Retailing*. Ph.D. thesis, Heinrich-Heine-Universität.
- Macqueen, J. 1967. Some Methods for Classification and Analysis of Multivariate Observations. *Pages 281–297 of: Berkeley Symp. on Math. Statist. and Prob.*
- Martin, Simon. 2024. Market transparency and consumer search—Evidence from the German retail gasoline market. *The RAND Journal of Economics*, **55**(4), 573–602.
- McLachlan, Geoffrey J., Lee, Sharon X., & Rathnayake, Suren I. 2019. Finite mixture models. *Annual Review of Statistics and Its Application*, **6**(1988), 355–378.
- Metropolis, Nicholas, Rosenbluth, Arianna W., Rosenbluth, Marshall N., Teller, Augusta H., & Teller, Edward. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092.

- Raftery, Adrian E. 1995. Bayesian Model Selection in Social Research. *Sociological Methodology*, **1995**(25), 111–163.
- Scheffman, David T., & Spiller, Pablo T. 1987. Geographic Market Definition under the U. S. Department of Justice Merger Guidelines. *The Journal of Law & Economics*, **30**(1), 123–147.
- Stigler, G ; Sherwin, R. 1985. The Extent of the Market. *Journal of Law and Economics*, **28**(3), 555–585.
- U.S. Department of Justice and Federal Trade Commission. 2023 (December). *2023 Merger Guidelines*. Issued by the U.S. Department of Justice and the Federal Trade Commission.
- Yang, Yan. 2018. *A New Solution to Market Definition: An Approach Based on Multi-dimensional Substitutability Statistics*. Ph.D. thesis, Washington University.
- Zwanziger, Jack, Melnick, Glenn A., & Mann, Joyce M. 1990. Measures of hospital market structure: a review of the alternatives and a proposed approach. *Socio-Economic Planning Sciences*, **24**(2), 81–95.

Appendix A Firm-Based Definitions

For a Firm-Based Market, we define the dual definition.

Definition 5 (Firm-Based Market) *A Firm-Based Market is a subset of firms $\mathcal{J}_g \subseteq \mathcal{J}$ such that*

$$C_j^{-1} = C_{j'}^{-1} \quad \forall j, j' \in \mathcal{J}_g$$

and no strict superset of $\mathcal{J}_g$ satisfies this property.

This definition states that a market is composed of all firms that have the same exposure set. There might, however, be firms in different markets that have the same consumers in their exposure sets.

The dual definition for the Firm-Based Market is:

Definition 6 (Firm-Based Market) *A Firm-Based Market is a subset of firms $\mathcal{J}_g \subseteq \mathcal{J}$ such that for any partition $B \subset \mathcal{J}_g$,*

$$\bigcup_{j \in B} \mathcal{C}_j^{-1} \cap \bigcup_{j \in \mathcal{J}_g \setminus B} \mathcal{C}_j^{-1} \neq \emptyset,$$

and no strict superset of $\mathcal{J}_g$ satisfies this property.

The dual definition for the *Probabilistic Firm-Based Market*:

Definition 7 (Probabilistic Firm-Based Market) *A Probabilistic Firm-Based Market is a subset of firms $\mathcal{J}_g \subseteq \mathcal{J}$ such that*

$$\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j'} \quad \forall j, j' \in \mathcal{J}_g,$$

and no strict superset of $\mathcal{J}_g$ satisfies this property.

Appendix B Model

The data consists of pairs $\{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \mathbb{R}^2$ is the geographical position of the house, $y_i \in \{1, \dots, M\}$ is the real estate agent that sold house i , observe that there are N houses and M different real estate agents. The unobserved variable is $\{z_i\}_{i=1}^N$, $z_i \in \{1, \dots, K\}$. z_i is the market (or cluster)

assignment for house i . Furthermore we have the following random variables

$$\begin{aligned} &\{\mu_j\}_{j=1}^K, \mu_j \in \mathbb{R}^2, \\ &\{\Sigma_j\}_{j=1}^K, \Sigma_j \in \text{PD}_2, \\ &\Theta \in \mathbb{R}^{K \times M}, \text{ s.t. } \theta_{jl} \geq 0, \sum_{j=1}^K \theta_{jl} = 1. \end{aligned}$$

μ_j is the center values matrix for cluster j , Σ_j is the covariance matrix for cluster j and θ_{jl} is the probability that a random house in market j is sold by agent l . Furthermore we have the non-stochastic hyperpriors

$$\begin{aligned} &\mu_0, \lambda_0, \nu_0, \Psi_0, \quad \text{hyperpriors to } \mu_j, \Sigma_j, \\ &\alpha, \quad \text{hyperpriors to } \theta_j. \end{aligned}$$

We impose the following

$$\begin{aligned} &x_i \perp y_i | z_i, \\ &(x_i, y_i) \perp (x_j, y_j) | z_i, z_j, \Theta, \\ &p(z_i | y_i, \Theta) = \theta_{z_i, y_i}, \\ &x_i | z_i, \mu, \Sigma \sim \text{N}(\mu_{z_i}, \Sigma_{z_i}), \\ &y_i \sim \text{U}(\{1, \dots, M\}), \\ &\mu_j, \Sigma_j \sim \text{NIW}(\mu_0, \lambda_0, \nu_0, \Psi_0), \\ &\theta_l \sim \text{Dir}(\alpha). \end{aligned}$$

The joint distribution can then be written as follows, shorthand notation $x = \{x_1, \dots, x_N\}$, $y = \{y_1, \dots, y_N\}$, $\mu = \{\mu_1, \dots, \mu_K\}$ etc,

$$p(x, y, z, \mu, \Sigma, \Theta) = \prod_{i=1}^N p(x_i|z_i, \mu, \Sigma) p(z_i|y_i, \Theta) p(y_i) \prod_{j=1}^K p(\mu_j|\Sigma_j) p(\Sigma_j) \prod_{l=1}^M p(\theta_{jl})$$

$$\prod_{i=1}^N N(x_i|\mu_{z_i}, \Sigma_{z_i}) \theta_{z_i, y_i} \frac{1}{M} \prod_{j=1}^K N(\mu_j|\mu_0, \frac{1}{\lambda_0} \Sigma_j) \mathcal{W}^{-1}(\Sigma_j|\nu_0, \Psi_0) \prod_{l=1}^M \frac{1}{B(\alpha)} \theta_{jl}^{\alpha_l - 1}.$$

Now the goal is to find the posterior $p(z|x, y)$. Since this posterior is intractable we find a way to sample from the posterior instead. Here we present a Gibbs sampler, a way to sample a realization of a Markov chain with $p(z, \mu, \Sigma, \Theta|x, y)$ as its invariant distribution. To implement the Gibbs sampler we need to be able to sample from the following distributions

$$p(z|x, y, \mu, \Sigma, \Theta),$$

$$p(\Sigma|x, y, z, \mu, \Theta),$$

$$p(\mu|x, y, z, \Sigma, \Theta),$$

$$p(\Theta|x, y, z, \mu, \Sigma).$$

We start with $p(\Sigma|x, y, z, \mu, \Theta)$, this is a standard conjugate prior so

$$\Sigma_j|x, y, z, \mu, \Theta = \Sigma_j|x, z \sim \mathcal{W}^{-1}(\nu'_j, \Psi'_j),$$

$$n_j = \sum_{i=1}^N 1_{z_i=j},$$

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^N x_i 1_{z_i=j},$$

$$\nu'_j = \nu_0 + n_j,$$

$$\Psi'_j = \Psi_0 + \sum_{i=1}^N (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T 1_{z_i=j} + \frac{\lambda n_j}{\lambda_0 + n_j} (\bar{x}_j - \mu_0)(\bar{x}_j - \mu_0)^T.$$

In the same way we get the posterior for μ

$$\begin{aligned}\mu_j|x, y, \Sigma_j, \Theta &= \mu_j|x, z\Sigma_j \sim N(\mu'_j, \Sigma_j), \\ \mu'_j &= \frac{\lambda_0\mu_0 + n_j\bar{x}_j}{\lambda_0 + n_j}.\end{aligned}$$

The posterior for Θ is also standard conjugate prior calculations

$$p(\theta_j|x, y, z, \mu, \Sigma) = p(\theta_j|y, z) \propto p(z|\theta_j, y) \propto \prod_{i=1}^N \theta_{j,y_i} 1_{z_i=j} \prod_{l=1}^M \theta_{jl}^{\alpha_l-1}$$

so we conclude that $\theta_j|y, z$ is $\text{Dir}(\alpha'_j)$ distributed, where

$$\alpha'_{jl} = \alpha_l + \sum_{i=1}^n 1_{y_i=l} 1_{z_i=j}.$$

Finally we have the distribution for z .

$$p(z_i|x_i, y_i, \mu, \Sigma, \Theta) \propto p(x_i|z_i, \mu, \Sigma)p(z_i|y_i, \Theta) = N(x_i|\mu_{z_j}, \Sigma_{z_j})\theta_{z_i y_i},$$

so the posterior conditional distribution for z_i is categorical distributed with probability

$$P(z_i = j) = \frac{N(x_i|\mu_j, \Sigma_j)\theta_{jy_i}}{\sum_{j'=1}^K N(x_i|\mu'_{j'}, \Sigma'_{j'})\theta_{j'y_i}}.$$

Now the Gibbs sampler can be implemented as follows.

Input: $x, y, \nu_0, \mu_0, \lambda_0, \Psi_0, \alpha$

Result: Samples z_i^t from the posterior distribution;

Make an initial guess for z_i^0 , K-Means for example;

for $t=1 \dots T$ **do**

Sample $\Sigma^t \sim p(\Sigma|x, z^{t-1})$;

Sample $\mu^t \sim p(\mu|x, z^{t-1}, \Sigma^t)$;

Sample $\Theta^t \sim p(\Theta|y, z^{t-1})$;

Sample $z^t \sim p(z|x, y, \mu^t, \Sigma^t, \Theta^t)$;

end

return $\{z^t\}_{t=burn}^T$

Gibbs sampler for the posterior distribution of cluster assignment

The burn in period is used because unless the original guess is drawn from the invariant distribution there will be a period until the samples are approximately from the stationary distribution.

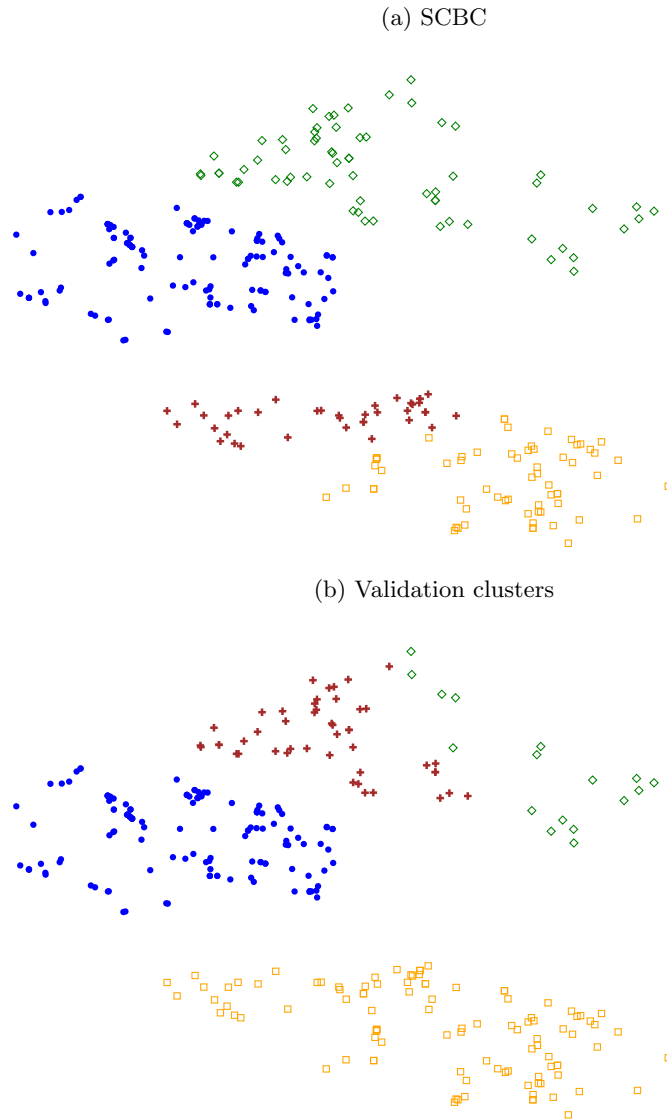
Appendix C Stockholm Real Estate

To illustrate further, we plot the clusters as scatter graphs. In Figure 10 we present the results of SCBC and compare it with the validation clusters. In the validation graph, we see that the southern cluster is on a separate island. The western cluster is also on an island, but because the water barrier is narrow, this is less visually obvious. The SCBC algorithm generates almost the same boundaries as the validation clusters, with the difference that the top clusters are merged and the bottom one is split into two.

In Figure 11 we plot the difference between SCBC and K-means. In this scatter plot, they look very similar. There are differences, but the scatter plot alone does not indicate which method performs better. We therefore place a

background of a map to see the relationship to geography.

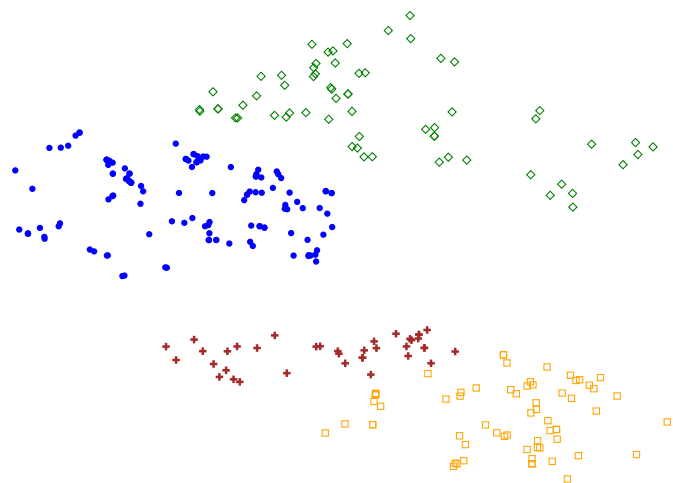
Figure 10: The clustering by SCBC with a DGMM compared to the listing areas



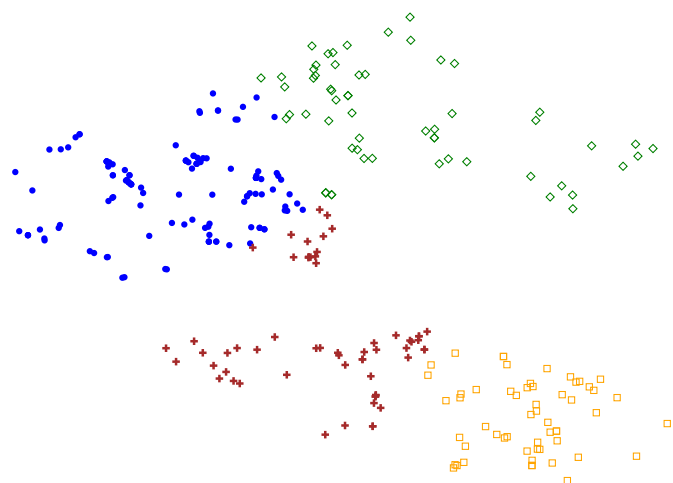
Notes: Comparing SCBC and the validation set. The different shapes correspond to the different classification into the clusters. The validation set is the listing areas.

Figure 11: The clustering by SCBC compared to K-means

(a) SCBC clusters



(b) K-means clusters



Notes: Comparing SCBC and K-means. The different shapes correspond to the different classification into the clusters.