Clustering with implicit constraints: A novel approach to housing market segmentation

Article *in* Transactions in GIS · December 2021

DOI: 10.1111/tgis.12878

6 authors, including:

Xiaoqi Zhang
Chinese Academy of Social Sciences
25 PUBLICATIONS **61** CITATIONS
SEE PROFILE

Yanqiao Zheng
University at Buffalo, The State University of New York
19 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Xinyue ye
Texas A&M University
329 PUBLICATIONS **5,221** CITATIONS
SEE PROFILE

Qiong Peng
University of Maryland, College Park
9 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Clustering with implicit constraints: A novel approach to housing market segmentation

Abstract

Constrained clustering has been widely studied and outperforms both the traditional unsupervised clustering and experience-oriented approaches. However, the existing literature on constrained clustering concentrates on spatially explicit constraints, while many constraints in the housing market studies are implicit. Ignoring the implicit constraints will result in unreliable clustering results. This paper develops a novel framework of constrained clustering, which takes into accounts implicit constrains. Specifically, the research extends the classical greedy searching algorithm by adding one back-and-forth searching step, efficiently coping with the order sensitivity. Via evaluating on both synthetic and real datasets, the proposed algorithm turns out outperforming the existing algorithms, even in the case that only the traditional pairwise constraints are provided. By applying to a concrete housing market segmentation problem, the proposed algorithm shows its power to accommodate the user-specified homogeneity criteria to extract hidden information of the underlying urban spatial structure.

Keywords: constrained clustering; implicit constraints; greedy algorithm; market segregation; urban spatial structure

1. Introduction

Housing market segmentation is a long-standing topic in the fields of spatial economics, urban planning, and economic geography(Bourassa et al., 2003; Goodman and Thibodeau, 1998; Galster, 1997; Goodman and Thibodeau, 2003; Guo et al., 2012; Helbich et al., 2013). However, it is challenging to identify a proper segmentation structure of housing market despite that persistent efforts have been made to the methodological development (Hepsen and Vatansever, 2012; Hwang and Thill, 2009; Islam and Asami, 2009; Miranda et al., 2017; Wu et al., 2018). The methods on housing market segmentation can be categorized into two types: experience-oriented and data-driven. Experience-oriented methods generate seg-

mentation directly from prior knowledge such as directly taking the segmentation generated by administrative boundaries or consulting real estate agents. Such methods are popular in practice because of the simplicity (Goodman and Thibodeau, 1998, 2003; Helbich et al., 2013). But due to the subjectivity, segmentation by prior knowledge is more likely to mistakenly break up the potential social-economic connectivity between two submarkets, and/or unify two classes of housing units together that should haven been separated.

Data-oriented methods, e.g. clustering approaches, rely on the high-dimensional feature data sampled for housing units in a market. Segmentation based on detailed features tends to outperform the experience-oriented methods in capturing the ground truth of housing submarkets that are spontaneously determined by the unobserved social-economic equilibrium. However, the data-oriented approaches may not always generate the desirable segmentation results. First, the increasing sample size makes the segmentation results intractably complicated which lacks of interpretability for human being. Second, the number of features is always giant for housing studies, such as the amenity conditions and neighborhood attributes, and the segmentation result is subject to the choice of features used for clustering, which complicates the implementation of data-oriented methods.

The ideal approach that combines benefits of the aforementioned approaches is constrained clustering. The constrained clustering method can combine researchers' prior partial knowledge on the cluster structure into the data-oriented clustering process, which turns the clustering from a unsupervised learning task to a semi-supervised one. The semi-supervised learning process reduces the arbitrariness of experience-oriented approach and remedy to the over-complexity issue of purely data-oriented segmentation. However, there are some limitations in the existing literature of constrained clustering. It is often assumed that the background knowledge can be expressed explicitly via the classical binary constraints (must-link/cannot-link) as introduced in Davidson and Ravi (2005); Wagstaff et al. (2001), or the cluster size constraints discussed in Fisher et al. (2015); Zhu et al. (2010). The explicitly expressed constraints are not sufficient for the segmentation tasks in housing market, because for housing market segmentation, the submarket structure is ultimately the consequence of the integration of tons of individual purchase decisions. While for these decisions, there are too many dimensions across which households have to make trade-offs, blurring the line between submarkets and suggesting the lack of universally effective constraints. A more realistic scenario is that the proper constraints are always case-dependent and vary significantly

according to the change of the research target and the user-specified segmentation criteria. To accommodate the customized constraints in different research scenarios, the classical binary constraints and/or the cluster size constraints are far from enough. Therefore, the housing market segmentation calls for a more flexible framework to impose constraints to the base clustering algorithm, from which the clustering algorithm with implicit constraints arises. Without loss of generality, the implicit constraints are represented via an inequality system of the following form:

$$\begin{cases} f_1(C_i, \quad \mathbf{C}_{-i}) \leq 0; \\ \qquad\qquad \vdots \\ f_k(C_i, \quad \mathbf{C}_{-i}) \leq 0. \end{cases} \tag{1}$$

where f_i for $i = 1, \ldots, k$ is a set of known functions whose functional form depends on the specific application settings. C_i is an arbitrarily fixed cluster, the notation $\mathbf{C}_{-i}$ represents the collection of all clusters other than C_i. The expression in (1) implies that the function f_i could depend on both of all the instances within the given clusters and the instances in the complementary clusters.

It turns out that the canonical binary constraints and cluster size constraints can all be converted to the constraint (1) with sufficient amount of f_i's. In addition, the other frequently used segmentation criteria in literature (Bonhomme and Manresa, 2015; Karpatne et al., 2017; Zhang et al., 2020, 2019) that is not possible to be represented as the binary and/or cluster size constraints can also be translated into the form (1) via properly selected function f_is. In particular, the implicit constraints (1) can be set to reflect the economic rationality claimed in the urban economic theory, by which we can implement the economic-theory-guided market segmentation that makes the resulting submarket structure meaningful in both the economic and geographic sense. For instance, the classical urban economic theory (Alonso et al., 1964; Mills, 1972; Muth, 1969) predicts the positive transportation premium, which means given all others equal, the price of a housing unit should increase along with the increase in the transportation convenience of its neighborhood. However, empirical studies (Knight et at., 1979; Pun et al., 2007; Tan et al., 2019; Zhang et al., 2019) have found counter-evidences in the housing markets of many metropolitan areas. If insisting in the classical urban economic theory, the counter-evidences suggest that these markets should have subtly divided submarket structure so that the positive premium can be recovered

within each of these submarkets. To identify the proper submarket structure, in section 4 of this paper, we shall take the housing market of Hangzhou as an example and show how the positive premium requirement can be translated into a set of implicit constraints (1) so that the proposed algorithm in this paper can effectively solve it. It is remarkable that the classical similarity-based clustering algorithms, such as DBSCAN, OPTICS, CURE, BIRCH, REDCAP, Chameleon, K-Means etc.(Ankerst et al., 1999; Birant and Kut, 2007; Ester et al., 1996; Guha et al., 1998; Guo , 2008; Karypis et al., 1999; Liu et al., 2012; Zhang et al., 1996), are not helpful to generate satisfactory submarket structure, because the positive transportation premium constraint is intrinsically conflicting with the descending operation of the within-cluster difference requested commonly in similarity-based clustering algorithms. From this perspective, we believe the clustering technique with implicit constraints forms a powerful and irreplaceable tool for housing market studies.

Existing constrained clustering algorithms designed for binary and/or cluster size constraints may not be efficient to solve the clustering problem with implicit constraints (1). Because these algorithms are constraint-specific, relying heavily on the explicit expression of the constraints, which makes them not adjustable to handle the implicit constraints. Meanwhile, the potential nonlinearity in the function f_is incurs the sequential dependence issue (to be discussed in section 3), which is not taken into account by existing algorithms for constrained clustering. As a consequence, the clustering result would lose robustness and become sensitive to the order of data instances once if implicit constraints were introduced. To fill the gap, this paper develops a new greedy algorithm that is sufficiently flexible, its implementation does not require the constraints to be explicitly expressed, therefore can solve the clustering problem with the general form of constraints (1). At the mean time, the algorithm will give full respect to a major cause to the sequential dependence and help address its induced order sensitivity issue.

We proceed as follows. The next section briefly reviews the existing clustering methods applied to housing market segmentation and the well-developed algorithm for constrained clustering issues. The following section presents a formal description of the greedy algorithm proposed in this paper and validates it through applying the algorithm to a variety of datasets. We then apply our algorithm to a real data example of housing market segmentation. The final section replicates the key findings and offers implications.

2. Literature Review

2.1. Spatial heterogeneity in housing market and clustering segmentation

Spatial heterogeneity, namely non-stationary dynamics across space, refers here to spatial variations in housing prices and household preferences (Tang and Yiu, 2010; Wu and Sharma, 2012; Yao and Fotheringham, 2016; Yu et al., 2007). Existence of spatial heterogeneity makes the OLS-based hedonic price model (HPM) invalid, because of its stringent assumptions that the coefficients must be constant over all places, and its neglect of spatial effects (Kiefer, 2011; McGreal and Paloma , 2013; Tang and Yiu, 2010; Tse, 2002).

It is notable that spatial heterogeneity has close connection to structural features of urban area (Plaut and Plaut, 1998; Wang and Liu, 2013; Wang et al., 2016; Wang, 2017). For instance, the housing market of a poly-centric metropolis is usually composed of several disequilibrium submarkets (Bourassa et al., 1997; Feng et al, 2014; Goodman and Thibodeau, 1998; Islam and Asami, 2009; Wu and Sharma, 2012). Spatial heterogeneity then comes from different determinants of housing prices in various submarkets (Bohman and Nilsson, 2016; Goodman and Thibodeau, 1998, 2003; Ottensmann et al., 2008).

Because the division of housing submarkets can help delineate the spatial heterogeneity behind the entire housing market of a city and improve the performance of housing price models (Goodman and Thibodeau, 1998, 2003), it becomes demanding to have an efficient method to identify housing submarkets from real data. Clustering, as a classical unsupervised learning task, is devoted to delineate data instances into a collection of disjoint subsets (partition) without reliance on any prior knowledge regarding the desired partition. Therefore, it is perfectly applicable to the housing market division problem, especially when there are no clues for the submarket structure. In literature of housing studies, many classical clustering algorithms have been applied to segment housing submarkets, such as the K-Means and its variants(Helbich et al., 2013; Wu et al., 2018; Yu et al., 2011), density-based clustering including DBSCAN(Ester et al., 1996; Guo et al., 2012) and its variants, hierarchical clustering (Estivill-Castro and Lee, 2002; Hepsen and Vatansever, 2012; Soaita and Dewilde, 2019), fuzzy clustering (Gabrielli et al., 2017; Hwang and Thill, 2009), clustered linear regression (Bonhomme and Manresa, 2015), clustering based on self-organization map (Fernando et al, 2005) and so on.

To account for the economic meaning of the divided submarkets, in most applications, not

only the spatial attributes, but the economic features of housing units, such as the neighborhood amenities, are also frequently utilized for clustering operation (Guo et al., 2012; Soaita and Dewilde, 2019; Wu et al., 2018; Wu et al., 2020). It is also found that in housing applications, compared to spatial attribute, the non-spatial attribute of housing units often plays a different role in determining the submarket structure. Therefore, the hybrid clustering approaches are developed. In Deng et al. (2011); Liu et al. (2019), the density-based clustering criteria is applied to the spatial dimension while the K-means-styled criteria is applied to the non-spatial dimension. By extending the classical DBSCAN algorithm, Birant and Kut (2007) develops the ST-DBSCAN algorithm that applies different radius parameters to the spatial and non-spatial dimension of the attribute vector. Although the hybrid approaches turn out performing better in the identifying housing submarkets, these novel approaches share the common feature with the classical approaches, such as DBSCAN and K-means, that the membership of data instance is always assigned in the direction that guarantees the within-cluster difference constantly descending. Despite the difference in the similarity measures of data instances, the descending in the within-cluster difference suggests that all the aforementioned approaches are similarity-based, which requires the latent assumption that the homogeneity within a cluster means the little difference among data instances in their attribute value. Being similarity-based also implies that the aforementioned approaches may not perform well for the clustering task that the within-cluster homogeneity does not rely on the similarity between data instances. One example of the non-similarity-based clustering task is discussed in section 4 of this paper where the within-cluster homogeneity is defined via the positive transportation premium constraint. By this constraint, every cluster of the homogeneous data instances should share the identical co-variation trend on some of their attribute variables, where the identical co-variation trend requirement naturally rejects the small within-cluster difference. The non-similarity-based clustering task also suggests the necessity to apply constrained clustering into housing market segmentation.

2.2. Constrained clustering

It is notable that the prior knowledge on the submarket structure is not always fully absent in practice. But on the other hand, a complete division of submarkets cannot be recovered only from the prior knowledge. In this case, the prior knowledge is only partial knowledge, which cannot be utilized independently, but is beneficial if combined together

with clustering algorithms. Since the required input structure of traditional clustering algorithms does not support them to take advantage of prior knowledge, a new branch of literature is emerging that facilitates the inclusion of background knowledge (Basu et al., 2004, 2008; Davidson and Ravi, 2005; Diego et al., 2017; Fisher et al., 2015; Hong and Kwong, 2008; Wagstaff et al., 2001). As a mixture of supervised and unsupervised learning techniques, constrained clustering studies emerge with a growing popularity.

The background knowledge is expressed as a set of constraint conditions imposed either on the instance level, such as the classical must-link and cannot-link constraints introduced in Davidson and Ravi (2005); Wagstaff et al. (2001), or on the cluster level, such as the requirement on the minimal number of members within every cluster (Fisher et al., 2015; Zhu et al., 2010). Algorithms to solve the clustering problem with these constraints have been well developed, such as the classical COP-K-Means (Wagstaff et al., 2001), PCK-Means (Basu et al., 2004) and the constrained programming approaches (Dao et al., 2017) for the pairwise constraints, the integer-linear-programming-based algorithm (Tang et al., 2020; Zhu et al., 2010) for size constraints, and many extensions to them (Davidson and Ravi, 2005; Diaz-Valenzuela et al., 2016; Fisher et al., 2015; Gancarski et al., 2020; Hong and Kwong, 2008; Karpatne et al., 2017; Le et al., 2018; Randel et al., 2018; Yu et al., 2015). These algorithms are mainly developed from the computing perspective, focusing on the algorithm efficiency. However, many housing problems can only be converted to a clustering task with implicit constraints, such as the housing market segmentation discussed in Zhang et al. (2020, 2019), for which the implicit constraints (1) cannot reduce to the familiar pairwise constraints or the size constraints. To this end, novel algorithms are needed.

3. Method

3.1. Sequential dependence and the failure of existing algorithms

Implicit constraints entail sequential dependence which makes clustering result hypersensitive to the order of assigning data instance into clusters (Hong and Kwong, 2008). Consider the following two types of sequential dependence:

Missing-inclusion: The instance d_i can be inserted into the cluster C_j without violating (1) only if a set of the other instances M_{d_i,C_j} have already been included in C_j and the set M_{d_i,C_j} depends on both the given instance d_i and the instances already included in C_j.

Including-exclusion: The instance $d_i \in C_j$ excludes the possibility that a set of the other instances E_{d_i,C_j} can be inserted in C_j, i.e. if d_i is not inserted in C_j, all instances in E_{d_i,C_j} can be put into C_j without violating (1).

It is easy to check that the including-exclusion type of sequential dependence can be induced by the classical cannot-link constraint (Davidson and Ravi, 2005; Wagstaff et al., 2001). In contrast to the including-exclusion, the missing-inclusion type of sequential dependence is more complicate, as it cannot arise from the usual pairwise must-link or cannot-link constraints nor from the size constraints. But the missing-inclusion issue does appear when implicit constraints (1) are introduced. As an illustration, we present a synthetic example which reveals how the missing-inclusion[1] type of sequential dependence arises from the constraints of the form (1) and how the existing algorithms suffer from the order sensitivity due to the existence of missing inclusions.

Consider the following graphic example in Fig. 1. Fig. 1a represents the set of data instances, where the red points are uniformly drawn from the triangle with three vertices at $(0, -\frac{2}{3})$, $(0, \frac{1}{3})$ and $(\frac{2}{3}, 1)$. The blue points are uniformly drawn from the disk with radius 0.5 and centred at $(0.9, -0.5)$. The ground truth consists of two clusters that are represented by the red-point set and blue-point set respectively. It turns out that both of the two clusters satisfies the following constraints:

$$\begin{cases} \sum_{i \in C_j} x_{i,2} \leq 0 \\ -\min_{i \in C_j, i' \in C_{-j}} d(x_i, x_{i'}) + c \leq 0 \end{cases} \tag{2}$$

where x_i is the two dimensional feature vector corresponding to the ith instance, $x_{i,2}$ is the second coordinate of x_i, C_j is the jth cluster with $j = 1, 2$, C_{-j} follows the definition in (1), d is the Euclidean distance on the plane. c is a constant equal to the minimal distance

[1]We do not focus on the including-exclusion type of sequential dependence here. It is because that in many settings, the including exclusion is simply a consequence of the existence of missing inclusions. Once if all missing inclusions have been added, there won't be including-exclusion issue any more. The example in the next section reveals this point. The other reason is that the missing-inclusion type of sequential dependence is relatively new to the literature of constrained clustering, to our best knowledge, there has not yet been any existing algorithm that attempts to handle the order sensitivity brought by missing inclusions. So, we hope our work could provide a starting point for algorithm design in this direction. The concentration on the missing-inclusion does not mean that the including-exclusion type of sequential dependence is not important, the including-exclusion is critical as well but it has been well discussed in many existing papers regarding the algorithm design to handle cannot-link constraints (Hong and Kwong, 2008), so we won't take it as priority and leave it for future studies.

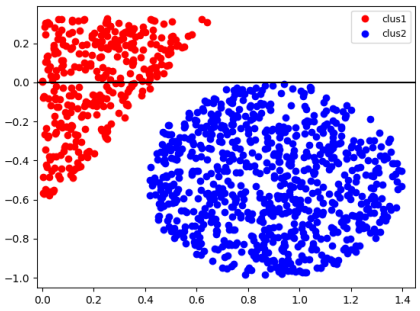
between all pairs of instances in the two clusters, in the current case, it is 0.21.

Suppose the instances in the cluster of red points are always ordered in prior to the instances in the blue-point cluster, and within each cluster, the instances are ordered purely randomly. Given this default order, we apply the COP-K-Means algorithm introduced in Wagstaff et al. (2001), the convergent clustering result is shown in Fig. 1b-1d.

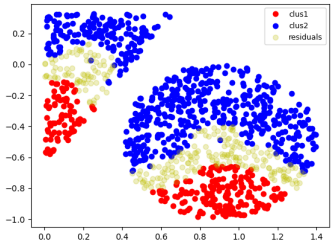
Fig. 1b shows the clusters resulting from the first iteration of instance assignment (initialization) before the center update. Fig. 1c and 1d display the clusters when the algorithm terminates, the two figures result from different default orders. From Fig. 1c and 1d, the two clusters generated by COP-K-Means are totally messed up and there are many instances that cannot be grouped into either cluster and be retained in the residual set, meanwhile the resulting clusters are sensitive to the choice of the default order. This result is a direct consequence to the missing-inclusion type of sequential dependence. The missing-inclusion is caused by the inconsistency between the default order and the first line constraint in (2). In fact, within the red-point cluster, all instances are purely randomly ordered, then the instances that have positive value in their second coordinate, denoted as PR, do not have to be ordered after those instances with negative second coordinate. Since the COP-K-Means always follows the default order to decide the assignment of every instance, it is very likely that an instance in PR is picked so early that once if it is added into the red-point cluster, the first-line constraint in (2) gets broken. Consequently, this instance has to be assigned either to the residual set or to the blue-point cluster, in either case, the assignment is incorrect, as revealed by the initialization in Fig. 1b.

When the ratio of the incorrectly assigned instances is large at the early stage, the miss-assignment trend becomes irreversible, because the red-point instances assigned to the blue-point cluster would move the center of the blue-point cluster away from its correct position. After multiple iterations, the centres for both clusters are relocated and the iteration will be locked into an incorrect status until termination. Finally the COP-K-Means poorly converges to a weird clustering result, as shown in Fig. 1c and 1d.

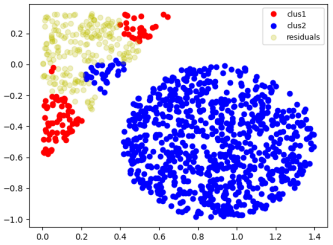
Fig. 1f and 1g shows the first-iteration and the convergent clusters generated by the algorithm developed in this paper, in which a back-and-forth searching is utilized to avoid the distortion induced by the early-pick-up of the instances in PR. Consequently, both the initialization and the further iteration won't suffer from the missing-inclusion type of sequential dependence. The correct clusters is perfectly resumed.



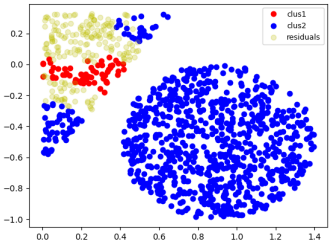
a. Sample data



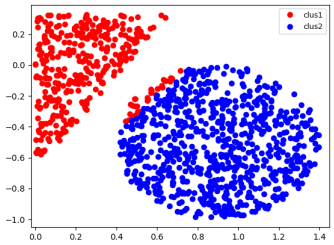
b. Initialization for COP-K-Means



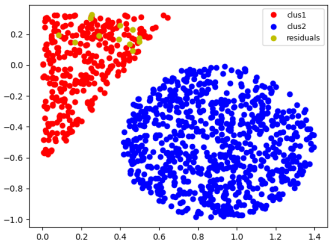
c. Absorbing state COP-K-Means



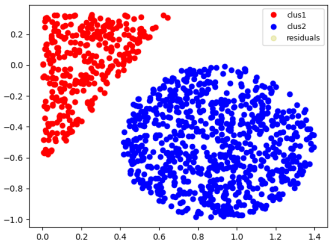
d. Absorbing state COP-K-Means



e. Unsupervised K-Means



f. Initialization for BFS-K-Means



g. Convergent BFS-K-Means

Figure 1: Clustering Example

Finally, Fig. 1e gives the clusters resulting from the unsupervised K-Means algorithm. Since the constraint (2) is not included at all, the cluster membership is completely determined by the distance between every instance and every cluster center. As shown in Fig. 1e, there exists a part of boundary points that should be contained in the blue-point cluster by the ground truth in Fig. 1a. But they are closer to the center of the red-point cluster in terms of the standard Euclidean distance, therefore, they are assigned to the red-point cluster by the unsupervised K-Means. However, grouping these boundary points into the red-point cluster would violate the second-line constraint in (2) that requires instances within every cluster have to be distributed tightly enough, the miss-assigned boundary points induce sparse distribution and cause contradiction to the constraint (2). This fact verifies the

necessity to include constraints in the clustering operation.

Remark 3.1. The graphic example shown in Fig. 1 reveals how the missing-inclusion type of sequential dependence can arise from the a set of implicit constraints, and how the missing inclusion can cause failure for the classical constrained clustering algorithm and the unsupervised K-Means algorithm. In addition to the missing-inclusion issue, the example 1 also sheds light to the including-exclusion issue. In fact, if a red point above the zero line in Fig. 1a is added into the red-point cluster at the time that after add-in, the first-line constraint in (2) is binding, then the newly added red-point will hold up the feasibility of adding any more red points above the zero line until a sufficient amount of red points below the zero line being added. Apparently, in this case, the newly added red point generates the including-exclusion type of sequential dependence, while this dependence issue can also be considered as with the missing-inclusion type. This is because the infeasibility of adding more red points with positive second coordinate is also a consequence that there are still many red points below the zero line that have not yet been included. Therefore, by properly arranging the inserting order, this kind of including-exclusion issue can be identified as a missing-inclusion issue and be effectively resolved by our algorithm as shown in Fig. 1g.

3.2. Algorithm design

A new greedy algorithm is developed here to resolve the missing-inclusion type of sequential dependence induced by the implicit constraints (1). Without loss of generality, we assume the baseline clustering problem is a K-Means clustering problem, it has the following baseline objective function:

$$\sum_{k=1}^{K} \sum_{i \in C_k} (D_{i,x} - \overline{D}_k)^2 \tag{3}$$

where C_k is the set of indices that specifies the instances belonging to the kth cluster. D_i is feature vector for the ith instance in the entire data set D, $D_{i,x}$ is a subvector of D that is the projection of D_i to the coordinates indexed by x which can be interpreted as the dimensions for clustering, $\overline{D}_k = \frac{1}{|C_k|} \sum_{i \in C_k} D_{i,x}$ is the center of C_k with respect to the clustering dimension x. Without loss of generality, we let the deviation from the center be calculated from some sub-coordinate x of the entire feature vector, because the latitude and longitude of housing units can be associated with contextual attributes of housing market.

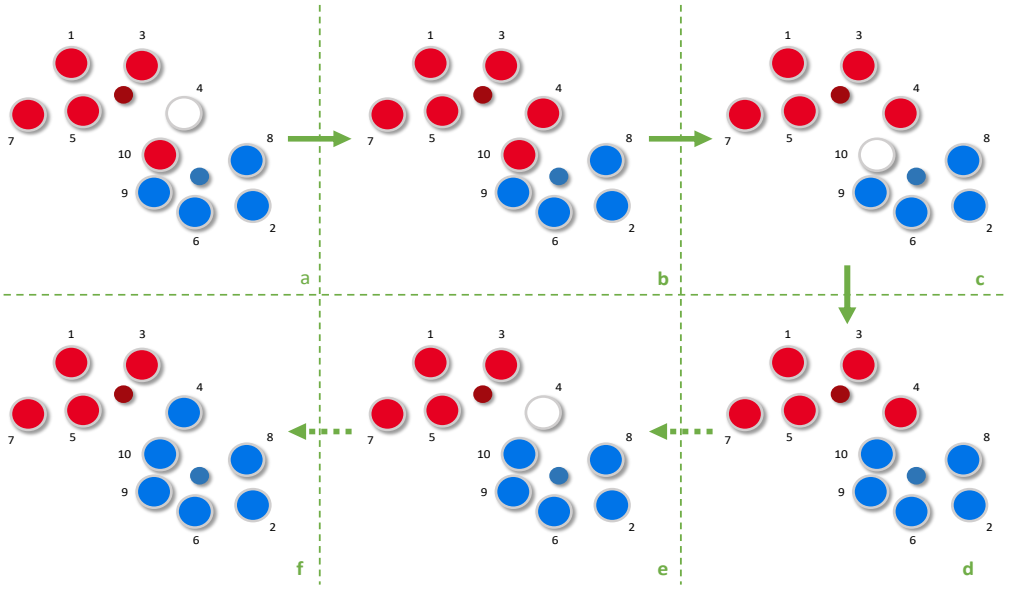


Figure 2: Illustration of the working mechanism for the trace-back M-step

As revealed by Fig. 1, the main issue behind missing inclusion is that sometimes the infeasible assignment of a given instance A is only a temporal phenomena, because it is mistakenly ordered behind some "wrong" instances in the queue but ahead of those "right" instances that guarantee the feasible assignment of A. After the "right" instances are assigned, A has been passed and won't be traced back any more. Therefore, the key to resolve the missing-inclusion issue is to save those mis-passed feasible instances. This fact inspires us to generalize the classical expectation-minimization (EM) algorithm for unconstrained clustering (3) by adding an iterative trace-back procedure to re-check the potentially feasible assignment of those mis-passed instances. The classical EM algorithm iterates the expectation (E-step) and minimization (M-step) operation until convergence. In the E-step, the algorithm computes the mean values $\overline{D}_k$, which turns out minimizing (3) for fixed membership relations of instances. In the M-step, instances are relocated among clusters so as to minimize (3) given the mean value $\overline{D}_k$s are fixed. Because missing-inclusion only impacts the effectiveness of the M-step, the iterative trace-back procedure is added to the M-step, its working mechanism is demonstrated in Fig. 2.

In Fig. 2, the ground truth includes 10 data instances that are separated into two clusters (as shown in Fig.2f). In the ground truth, the constraint is that for every instance A, the distance between A and the nearest instance to A within the same cluster must be less than the distance between A and the nearest instance to A outside the cluster that A belongs

to. Given the prior centers (represented as the two dark circles which may not be the true centers while can be viewed as the centers yielding from the previous E-step), Fig. 2a - Fig. 2d demonstrate how the M-step in existing algorithms, such as PCK-Means, proceeds to relocate the instances following the default order (indicated by the number label next to each data instance in sub-plots of Fig. 2). Given the previous-run assignment of instances represented by blue/red colors, Fig 2a shows that the M-step in the current run proceeds to the decision of assigning the 4th instance, which, due to the constraint, can only be assigned to the red cluster as shown in Fig.2b. Then, the M-step moves forward to the assignment decision of the last instance in Fig. 2c, and assign that to the blue cluster in Fig.2d, which terminates the traditional M-step. Throughout Fig. 2a - Fig. 2d, the 4th instance induces the missing-inclusion type of sequential dependence as it cannot be correctly assigned until that the 10th instance is embedded into the blue-colored cluster. But after the 10th instance is assigned, the M-step procedure terminates which leaves the mis-assignment unfixed. To this end, we add a trace-back procedure which will be intrigued whenever a real relocation occurs. In the case of Fig. 2, after the 10th instance being relocated to blue cluster, the trace-back is activated and the M-step will be repeated again following the same execution order, which leads to the second-run assignment decision of the 4th instance in Fig. 2e by which that instance will ultimately be relocated to the correct cluster. Since after that, there won't be any more relocation, the trace-back procedure will terminate at the status of Fig. 2f, and the next-run E-step starts. Formally, the trace-back M-step can be summarized into the following algorithm TBM.

For constrained K-Means clustering, the order involved in the Algorithm TBM can be naturally identified with the ascending order of the distance difference $DM_{i,j} - DM_{i,j_i}$s, where C_{j_i} is the cluster that instance i belongs to in the prescribed partition. Given the trace-back M-step, the classical EM algorithm can be implemented, the resulting clusters are obtained when the algorithm TBM converges such that $\mathcal{C} = \mathcal{C}'$. It turns out that replacing the standard M-step with the trace-back M-step does not change of greedy property of the EM algorithm. Therefore, the convergence can be guaranteed due to the finiteness of the input data and the greedy property.

Trace-Back M (TBM)

Require:

The set of all instances D; Fixed partition of the set of instances, $\mathcal{C} = \{C_1, \ldots, C_K\}$;

The matrix DM recording the distance between every instance i and the center of every cluster C_j;

A prescribed ordered set $order = \{(i, j_i, j) : i = 1, \ldots, |D|;\ j = 1, \ldots, K;\ i \in C_{j_i}\}$.

Ensure:

A renewed partition $\mathcal{C}' = \{C'_1, \ldots, C'_K\}$.

Set DM as the distance matrix $s.t. DM_{ij} = (D_{i,x} - \overline{D}_j)^2$

Set $position = 0$

while $position < order.length$ **do**

  Set $item = order[position]$

  Set $origin = item[1]$, $des = item[2]$, $id = item[0]$

  if $DM_{id,des} < DM_{id,origin}$ **then**

    Set $C'_{origin} = C_{origin}/\{id\}$, $C'_{des} = C_{des} \cup \{id\}$

    Set $\mathcal{C}'' = \left(\{C'_{origin}, C'_{des}\} \cup \mathcal{C}'\right)/\{C_{origin}, C_{des}\}$

    ReSet $position+ = 1$

    if Constraints (1) hold for new C'_{origin} and C'_{des} **then**

      ReSet $\mathcal{C}' = \mathcal{C}''$

      ReSet $order = \{element \in order : element[0] \neq id\}$

      ReSet $position = 0$ **(Trace-Back activation)**

    end if

  end if

end while

return $\mathcal{C}'$

Algorithm for a partially feasible constraints

A latent assumption behind the algorithm TBM is that there exists at least one partition of the full set of data instances that is consistent with the constraints. When there does not exist such a feasible partition or we can not find it during the initialization stage, we can release the feasible partition requirement to that there exists a subset $D' \subset D$ such that

the K-fold partition $\mathcal{C}_{D'}$ is feasible for D' under constraints (1). Then, the complement set $R = D/D'$ can be thought of as a residual set. The algorithm for the existence of feasible partition can be extended easily to the more general cases through allowing soft constraints and augmenting the violation to some of the constraints into the objective function (Basu et al., 2004; Davidson and Ravi, 2005; Zhu et al., 2010). Formally, we can reformulate the objective function of constrained K-Means clustering with partially infeasible constraints as below:

$$\sum_{k=1}^{K} \sum_{i \in C_k} (D_{i,x} - \overline{D}_k)^2 + c \cdot |R| \tag{4}$$

where the term $c \cdot |R|$ is added to punish the number of instances that cause violation to the constraints (1), in which c is a large positive constant satisfying that $c > \max\{(D_{i,x} - D_{i',x})^2 : i, i' = 1, \ldots, |D|\}$. Such choice of c implies the violation to (1) is punished so seriously that it always becomes better as long as an instance in the residual set can be classified into some clusters no matter how distant is between the instance and the cluster center. If we consider the set of residuals as a special cluster and define the distance between every instance i and the cluster R through

$$DM_{i,R} = \begin{cases} c & if\, i \notin R \\ 0 & else. \end{cases} \tag{5}$$

, then the trace-back M-step in algorithm TBM is completely applicable to the constrained K-Means clustering problem with partially feasible constraints such that the clusters terminating the algorithm TBM under the distance specification (5) forms a local minimum of (4) under constraints (1).

Algorithm for infeasible constraints

In the real-world applications, we cannot exclude such an extreme case that the constraints are too harsh to guarantee the existence of a partially feasible partition. Then, we cannot find any subset $D' \subset D$ and a partition, $\mathcal{C}$, of D' such that $\mathcal{C}$ is feasible under constraints (1). But even in this extreme setting, it is still expected that we can find out a partition of D such that the constraint inequalities (1) are close to hold. Formally, we can always assume that there exists a K-fold partition $\mathcal{C}$ of D and a set of (non-negative) tolerance thresholds $\mathbf{c} = (c_{11}, \ldots, c_{1k}; \ldots; c_{K1}, \ldots, c_{Kk})$ such that the the following relaxed

version constraints (1) can be satisfied by $\mathcal{C}$ up to the tolerance $\mathbf{c}$:

$$\begin{cases} f_1(C_i, \quad \mathbf{C}_{-i}) \leq c_{i1}; \\ \qquad\qquad \vdots \\ f_k(C_i, \quad \mathbf{C}_{-i}) \leq c_{ik}. \end{cases} \tag{6}$$

With such an initialization $\mathcal{C}$, we hope to search for the best clustering result that solves a soft-constrained K-Means clustering problem, which is equivalent to the following unconstrained minimization problem:

$$\min_{\mathcal{C}=\{C_1,\ldots,C_K\},c_{li}\geq 0\forall i,l} \sum_{l=1}^{K}\sum_{i\in C_k}(D_{i,x}-\overline{D}_l)^2 + \sum_{l=1}^{K}\sum_{i=1}^{k}\max(f_i(C_l,\mathcal{C}_{-l})-c_{li},0) \tag{7}$$

In problem (7), the constraints (6) are augmented into the objective function through a punishment term. The augmentation is different from (4). in (4), the punishment is on the number of instances that violate the tough constraints (1). While due to the non-existence of a D' partially feasible partition, the punishment in (4) becomes trivial. In (7), a weaker punishment condition is applied which is on the summation of the degrees that every constraint condition is violated and the violation degree is measured by the value of the constraint function f_is.

By duality, the procedure to solve the softly constrained problem (7) can be treated as a process that gradually tightens the tolerance threshold $\mathbf{c}$ which would converge to a local optimal if no any single coordinate of $\mathbf{c}$ can be further compressed. Following this idea, the following modified TBM algorithm can be taken as the modified M-step to solve (7):

Soft Trace-Back M (Soft_TBM)

Require:

The set of all instances D; Fixed partition of the set of instances, $\mathcal{C} = \{C_1, \ldots, C_K\}$ and the associated violation degrees $\mathbf{c}$;

The matrix DM recording the distance between every instance i and the center of every cluster C_j;

A prescribed ordered set $order = \{(i, j_i, j) : i = 1, \ldots, |D|;\ j = 1, \ldots, K;\ i \in C_{j_i}\}$.

Ensure:

A renewed partition $\mathcal{C}' = \{C'_1, \ldots, C'_K\}$.

Set DM as the distance matrix $s.t. DM_{ij} = (D_{i,x} - \overline{D}_j)^2$

Set $position = 0$

while $position < order.length$ **do**

  Set $item = order[position]$

  Set $origin = item[1]$, $des = item[2]$, $id = item[0]$

  if $DM_{id,des} < DM_{id,origin}$ **then**

    Set $C'_{origin} = C_{origin}/\{id\}$, $C'_{des} = C_{des} \cup \{id\}$

    Set $\mathcal{C}'' = \left(\{C'_{origin}, C'_{des}\} \cup \mathcal{C}'\right)/\{C_{origin}, C_{des}\}$

    ReSet $position+ = 1$

    if $f_i(C'_l, \mathcal{C}''_{-l}) < c_{li},\ \forall i = 1, \ldots, k;\ \forall l = origin, des$ **then**

      ReSet $\mathcal{C}' = \mathcal{C}''$

      ReSet $order = \{element \in order : element[0] \neq id\}$

      ReSet $c_{li} = f_i(C'_l, \mathcal{C}''_{-l})$ for $i = 1, \ldots, k$, $l = origin, des$

      ReSet $position = 0$ **(Trace-Back activation)**

    end if

  end if

end while

return $\mathcal{C}'$

Remark 3.2. Notice that the algorithm $Soft\_TBM$ is an extension to the algorithm TBM in the sense that algorithm $Soft\_TBM$ reduces to TBM as long as the set of initial violation threshold vector $\mathbf{c}$ is 0. Despite that, the execution of algorithm $Soft\_TBM$ differs from

that of the algorithm TBM in that they need different initialization unless $\mathbf{c}$ is set to 0. In addition, when the initial tolerance threshold $\mathbf{c}$ is not 0, the algorithm $Soft\_TBM$ updates both the membership of every instance and the tolerance threshold simultaneously, but the algorithm TBM updates memberships only.

Since the choice of the algorithm TBM and $Soft\_TBM$ depends on the feasibility degree of the constraints, and in general it is not possible to exclude the fully infeasible constraints. In practice, we will firstly try algorithm TBM and generate random initialization up to N times. As long as a partial feasible partition is generated within the N trails, the algorithm TBM will be selected to implement further optimization. If there is no partial feasible partition available for all the N trails, the algorithm $Soft\_TBM$ will be executed with the initial threshold vector $\mathbf{c}$ initialized as the smallest evaluation of the constraint function during the N trails. In the following sections, we set the number $N = 100$. For the initialization, we generate it via an algorithm similar to the classical PCK-Means algorithm (Basu et al., 2004). The only difference between PCK-Means and our initialization algorithm is that the PCK-Means would return output until convergence. Our initialization algorithm follow the same iteration of PCK-Means but wont wait until converge, instead, it returns the output one-iteration ahead of convergence. Therefore, the resulting initialization can be considered as a random shock from the optimal solution of PCK-Means, we then focus on detecting whether our algorithm can further improve the optimal performance of PCK-Means.

3.3. Evaluation

To evaluate the algorithm, we apply it to two synthetic datasets generated from a simple version of the model (9) that arise from the application discussed in section 4 and the model discussed in Zhang et al. (2020). The particular form of the constraints for the two datasets follow (10) and the setup in Zhang et al. (2020). The algorithm is also applied to the breast-cancer dataset (https://archive.ics.uci.edu/ml/datasets/breast+cancer) from the UCI repository. The breast-cancer dataset is selected because it is a classical data set and applied to test the algorithm performance of a variety of constrained clustering algorithm (Basu et al., 2004; Davidson and Ravi, 2005; Zhu et al., 2010), therefore we follow theconvention and include it in this comparison analysis. The comparison is made among the algorithm developed in this paper, the PCK-Means algorithm proposed in Basu et al.

(2004) (with proper modifications made to accommodate the implicit constraint (1)) and the unsupervised K-Means algorithm in which no prior constraint is available.

For the UCI breast-cancer data, we follow the literature (Basu et al., 2004; Davidson and Ravi, 2005; Zhu et al., 2010) to consider binary constraints only, and randomly generate a number of must-link and cannot-link constraints from the ground-truth cluster structure. As the implicit constraints (1) include the pairwise constraints as special cases, the algorithm designed in this paper also applies to the pairwise constraint case. The breast-cancer dataset provides a chance to evaluate the performance of our algorithm in the reduced scenario that only pairwise constraints are involved.

The construction details of the two synthetic datasets are as the following.

SD1. The data consists of 500 instances and 4 dimensions. The data is partitioned into 6 clusters with the number of instances in every cluster are randomly decided as an integer greater than 2. Within each cluster C_i, the data instances are randomly generated as the following:

1. randomly generate a two dimensional random vector uniformly from the square $[0, 3] \times [0, 3]$, denoted as $\mu = (\mu_{i1}, \mu_{i2})$;
2. generate $|C_i|$ two-dimensional random vectors from the normal distribution $N(\mu, I_2)$ where I_2 is the two-dimensional identity matrix, denote D_{i1} as the $|C_i| \times 2$ matrix storing the random data;
3. generate a random number, β, from the uniform distribution on $[0, 1]$ and a $|C_i|$ two-dimensional random vectors from the normal distribution $N(0, I_2)$, denoted as D_{i2}, define $D'_{i2} = (\beta \cdot D_{i2,\cdot 1} + D_{i2,\cdot 1}, D_{i2,\cdot 1})$ where $D_{i2,\cdot j}$ is the jth column of matrix D_{i2};
4. construct the four-dimensional data matrix $SD1_i = (D_{i1}, D'_{i2})$ for cluster C_i by augmenting the two two-dimensional data matrices.

The full data matrix are constructed through concatenating all the $SD1_i$ together, denoted as $SD1$. The random data set $SD1$ is a synthetic mimic to the data structure in application section 4, the first two columns of $SD1$ are the clustering column which can be interpreted as the rescaled latitude and longitude of a location, the constraints for this example follow the form of (10).

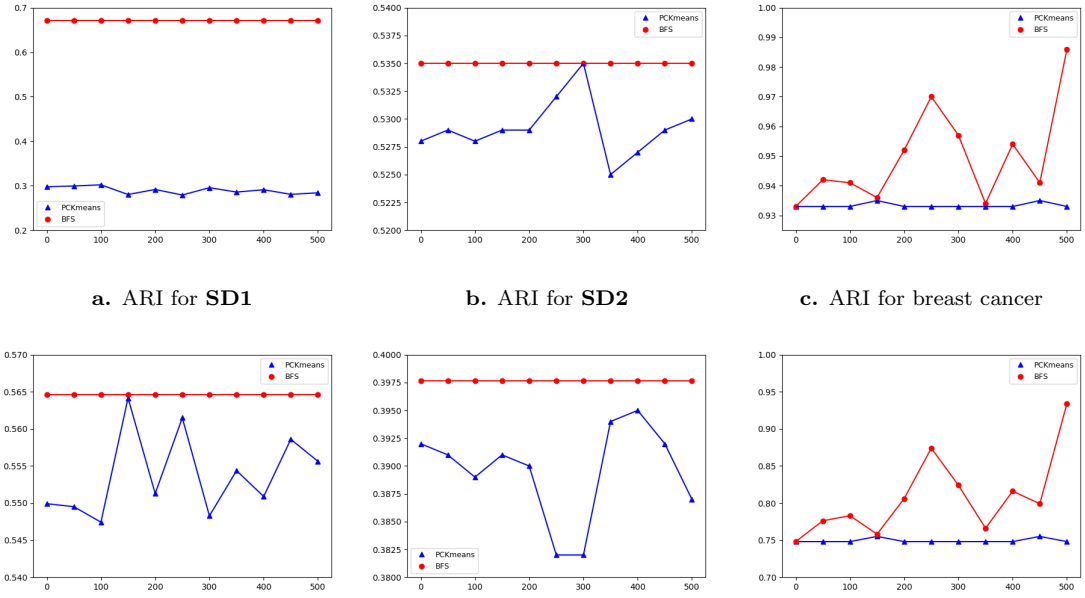
SD2. The data consists of 500 instances and 8 dimensions. The data is partitioned into 2 clusters with the number of instances in every cluster are randomly decided as an integer greater than 2. Within each cluster C_i, the first two columns are generated in exactly the same way as D_{i1} in previous paragraph and will be used as the clustering column, the remaining 6 columns are generated as the following:

1. for each $i = 1, 2$ generate $|C_i|$ random numbers valued in $\{-1, 1\}$ and denote the resulting vector as D_{i1};
2. generate a 2×2 random matrix M with each entry following uniform distribution on $[0, 1]$;
3. iteratively generate D_{ij} for $j = 2, \ldots, 6$ and $i = 1, 2$ as that i) for every index $k = 1, \ldots 500$, randomly select $k' \in \{1, \ldots, 500\}$, denote i_k as the index of the cluster that k is contained in and $D_{i_{k'}, j-1, k'}$ as the entry of the random vector $D_{i_{k'}, j-1}$ corresponding to the instance k', ii) generate an uniform random number s from $[0, 1]$ and if $s \leq M_{j_{k'}, j_k}$ assign $D_{i_{k'}, j-1, k'}$ to the entry $D_{i_k, j, k}$, otherwise, assign $-D_{i_{k'}, j-1, k'}$ to $D_{i_k, j, k}$;
4. the resulting D_{ij}s for $j = 1, \ldots, 6$ are appending together to form the remaining six column of $SD2_i$, the final data matrix $SD2$ are obtained from concatenating $SD2_1$ and $SD2_2$ toegther.

The $SD2$ is an synthetic analogue to the data structure discussed in Zhang et al. (2020).

For all the three datasets, when the PCK-Means is applied, we consider a sequence of increasing sets of constraints with the size increasing from 50 to 500 by 50, within each of the constraint set, the binary must-link and cannot-link constraints are randomly generated from the ground truth, by which we can detect the variation of the algorithm performance. The performance of clustering result is evaluated by the average value and the corresponding standard deviation of the adjusted rand index (ARI) and the Normalized mutual information (NMI), after 10 runs (Basu et al., 2004; Wagstaff et al., 2001).

The results of the experiments are shown in Fig. 3, where we plot the variation trend of the mean ARI and NMI resulting from the PCK-Means algorithm along with the increasing of the number of pairwise constraints, and the mean ARI and NMI generated by the algorithm



a. ARI for **SD1** **b.** ARI for **SD2** **c.** ARI for breast cancer

d. NMI for **SD1** **e.** NMI for **SD2** **f.** NMI for breast cancer

Figure 3: Fitting performance: BFS v.s. PCK-Means

proposed in this paper. As the standard deviation of ARI and NMI for all the algorithms is small with the scale of 10^{-4}, so we omit them in Fig. 3. In Fig. 3c and Fig. 3f, the algorithm proposed in this paper is also allowed to utilize the pairwise constraints, so the mean variation trend of the ARI and NMI for our algorithm can also change along with the number of pairwise constraints, which in the other two cases, as the number of constraints for our algorithm is fixed, the variation trend is trivial and represented as a horizontal line in the remaining four sub-figures. Also note that when the number of pairwise constraints is zeros the PCK-Means algorithm reduces to the unsupervised K-Means algorithm, therefore the mean ARI and NMI for PCK-Means at the zero point in all Fig. 3a-3f represents the result for the unsupervised K-Means.

From Fig. 3, our algorithm performs systematically better than both the unsupervised K-Means algorithm and the PCK-Means algorithm and this outperformance result won't be affected by the increasing number of pairwise constraints. Especially, even in the case that our algorithm and the PCK-Means take the same set of pairwise constraints, our algorithm still outperforms the PCK-Means significantly. According to Remark 3.2, the uniformly better performance yielded by our algorithm comes from the combination of the back-and-forth searching step in the algorithm design and the fact that our algorithm is initialized as a random shock to the optimal solution returned by PCK-Means. Without the random shock,

the sequential dependence might "cheat" the PCK-Means and lead it to terminate at the circumstance that an instance is only temporally suitable for a cluster while overall it is not really suitable. In PCK-Means, there is no mechanism to handle this temporal mismatch, while by adding the random shock, the back-and-forth searching and the re-evaluation procedure is activated, which grants more chances to re-consider the "appropriate" destination of every instance. Consequently, our algorithm can save the iteration from terminating at a relatively bad local solution, leading to a significantly better one, as shown in Fig. 3.

4. Application: A case study for the subway premium of housing price in Hangzhou, China

This section introduces a real data example in housing market studies. Through this example, we demonstrate how the implicit constraints (1) can be tailored to encode customized homogeneity criteria into the determination of housing submarket structure. In this example, the homogeneity criteria is derived from the classical urban economic theory which predicts that the housing price has positive transportation premium (Alonso et al., 1964; Mills, 1972; Muth, 1969). In the other words, given all others equal, the price of a housing unit should increase along with the increase in the transportation convenience of its neighborhood (Bajic, 1983; Feng et al, 2014; Islam and Asami, 2009; Li et al., 2011; Pels et al., 2010; Song et al., 2016; Wu and Sharma, 2012; Wu et al., 2017, 2016; Ye et al., 2007, 2018). It has been widely documented in empirical studies of housing price that the positive transportation premium hypothesis does not hold in housing markets of many metropolitan area, such as San Francisco, Hong Kong and Hangzhou (Knight et at., 1979; Pun et al., 2007; Tan et al., 2019; Zhang et al., 2019). As the positive transportation premium hypothesis holds only if all others equal, or equivalently, all the determinants of housing price but the transportation convenience must be unchanged. In most empirical studies, the condition that all others equal are hard to validate for the housing market associated with the entire metropolitan area, which suggests that submarket segmentation is needed in order to witness the positive transportation premium. Tan et al. (2019); Zhang et al. (2019) found supportive evidence for the positive transportation premium in the satellite area, which suggests a proper submarket division does help guarantee the "all others equal" condition. However, for the housing submarket in the central urban area, the evidence in Tan et al. (2019); Zhang et al. (2019) is still against the positive premium hypothesis. One way to interpret the result in Tan et al.

(2019); Zhang et al. (2019) is that the positive premium may still hold if we could identify a finer submarket structure for the housing market in the urban core region. Apparently, the desired submarket structure is hard to be figured out in an experienced-oriented way. But if we consider the inverse problem, which means we take the positive transportation premium as a homogeneity criteria and require that all housing units in a submarket (cluster) should be consistent with the positive premium hypothesis, the submarket structure might be easily identified with the assistance of the algorithm proposed in this paper. The example in this section would focus on identifying housing submarket structure of Hangzhou, China under the homogeneity criteria of positive transportation premium.

As commented in the introduction of this paper, there is not an universal criteria to define the homogeneity among housing units in a market, therefore whether a submarket structure is proper depends heavily on the validity of the homogeneity criteria taken by the clustering algorithm to generate that structure. Despite adopting the positive transportation premium as the homogeneity criteria in the following analysis, we do not attempt to claim how meaningful this criteria is, nor justify the reason that we select this criteria instead of the others. In fact, we just take the positive transportation premium as a user specified input, all our focus is on that given the input criteria, how it can be translated into a set of well-defined implicit constraints, and how our proposed algorithm can generate the desired submarket structure from these constraints.

Since the proper number of clusters is unknown in prior for all the real housing market data, we run the algorithm for all $K \in \{1, \ldots, 20\}$, the optimal cluster number K is determined by minimizing the objective function (3) for all different Ks. And for each K, we re-run the algorithm for 20 times and select the clustering result with the least (3) as the final result associated with the given K.

4.1. Problem reformulation

To identify submarkets that capture the spatial heterogeneity of correlation between housing price and subway system, the resulting segmentation must be both meaningful in spatial and economic sense. To guarantee the spatial meaning, we only adopt the two-dimensional spatial attributes, latitude and longitude, as the features for the K-Means objective function (3) which makes the resulting clusters spatially segregated. For the economic meaning, we encode it into the constraints by properly including the non-spatial attributes of housing

units, such as the area, building age, relative location and neighborhood amenities, into the mathematical formulation of the constraint. It is natural to separate the spatial and economic meaning via the separation between objective function and constraints, this operation distinguishes our method from the other modified clustering methods (Gabrielli et al., 2017; Wu et al., 2018) applied to housing market segmentation. In the later case, non-spatial features are utilized as a mixture with the spatial coordinate which makes the resulting submarkets spatially overlapped, lack of clear meaning in the spatial sense.

To set up proper constraints reflecting the positive transportation premium, we first quantify the effect of transportation convenience improvement incurred by high accessibility to subway system by following Zhang et al. (2019) to construct the metro index. Metro index encodes the information on both the local demand for express transportation and the improvement degree of transportation convenience brought by subway system. Formally, the metro index of a metro station is defined as the ratio of commuting time by no-subway routes to a set of major destinations in a city verses the commuting time by subway-prioritized routes, which can be expressed as the following:

$$metro\ \ index_i = \frac{1}{n}\sum_{j=1}^{n}\frac{route1_{ij}}{route2_{ij}} \tag{8}$$

where $route1_{ij}$, $route2_{ij}$ are the commuting time taken from station i to destination j by the optimal no-subway route and optimal subway-prioritized route, respectively. The choice of optimal route under both situations and the calculation of the corresponding commuting time are automated through the well-developed Baidu direction API (`https://lbsyun.baidu.com/index.php?title=webapi/direction-api-v2`). The choice of the number of major destinations and the destinations themselves are determined in a data-oriented manner by combining the hot-spot analysis (Hu et al., 2014) with the population density data, the details can be found in Zhang et al. (2019). From the definition (8), it is clear that the greater the metro index is, the higher commuting efficiency a metro station can induce to the residential units nearby. For a given housing unit, its metro index is defined as the metro index of its nearest metro station.

Based on metro index (8), the homogeneity criteri of positive transportation premium can be translated into a positivity constraints on the regression coefficient of the transportation convenience variable within the following hedonic housing price model (Can, 1990; Rosen,

1974),

$$P_i = \beta_0 + \beta_m \cdot X_i + \alpha \cdot Z_i + \mu_i \tag{9}$$

where P_i is unit price of the ith housing unit in the dataset, X_i is the ith observation of the metro index, Z_i is a vector of attribute variables associated with the ith housing unit which are treated as the control variables, μ_i is the residual, β_0, β_m and α are the intercept, coefficient for metro index and the coefficients for the control features, respectively.

Given the real housing market data, the hedonic model (9) together with the requirement of the positive transportation premium lead to the following particular form of the implicit constraints:

$$\hat{\beta}_{C,m} > 0, \;\; C \in \mathcal{C}_K \tag{10}$$

where $\mathcal{C}_K$ stands for a K-fold partition of the entire set of the housing price data and the feature variables of every housing unit. $\beta_{C,m}$ denotes the coefficient of metro index in the hedoic model (9) for the submarket corresponding to cluster C. $\hat{\beta}_{C,m}$ is the estimated value to $\beta_{C,m}$ by the classical ordinary least square (OLS) estimator. It turns out the OLS estimator depends in a highly nonlinear way on features of all housing units within a submarket, the non-linearity causes severely sequential dependence.

Constraint (10) requires that the estimated coefficient of metro index must be positive within every correctly identified submarket, reflecting that when all the other features hold the same, the housing price should react positively to the improvement of transportation convenience. Adding this constraint to the standard K-Means clustering can help generate the desired spatial segmentation result that is meaningful in both geography and economics, meanwhile satisfying the homogeneity criteria of positive transportation premium.

We also highlight that by the statistic theory of OLS estimator (Wooldridge , 2015), the key quantity $\hat{\beta}_{C,m}$ can be expressed as the following:

$$\hat{\beta}_{C,m} = I_m^\top (D_C^\top D_C)^{-1} D_C^\top P \tag{11}$$

where D_C is the $|C| \times p$ dimensional attribute matrix of all data instance included in cluster C, $|C|$ is the number of data instances assigned in C, p is the dimension of the non-spatial attribute vector (X, Z) of housing unit used as the explanatory variables in the hedonic model (9). Under this interpretation, for each data instance indexed by i in C, the ith row of D_C is then the row vector (X_i, Z_i). The vector P represent the collection of prices of housing units

contained in C, I_m is a row vector with dimension p in which only the coordinate associated with the attribute metro index takes value 1 while all the other coordinate take value 0. In fact, the complete matrix operation $(D_C^\top D_C)^{-1} D_C^\top P$ in (11) gives the OLS estimators for all the regression coefficients included in (9), while the row vector I_m helps select the one corresponding to $\beta_{C,m}$. By (11), we emphasize that although the constraint (10) focuses only on the positivity of the $\hat{\beta}_{C,m}$, it does not means the constraint put restrictions only on the two attributes, housing price and metro index. In the contrary, the value $\hat{\beta}_{C,m}$ depends on all the attribute variables, including both metro index and the other housing characteristics such as the building area, age, floor and so on. Therefore, constraint (10) gives a full respect to all the observable non-spatial housing characteristics, which helps control the potential heterogeneity arising these characteristics.

In addition, (11) suggests that $\hat{\beta}_{C,m}$ depends in a highly non-linear way on all of the housing units included in C, rather than merely on the pair-wise relation between some pair of housing units in C. The collective dependence of constraint (10) on all housing units in C incurs the sequential dependence as discussed in section 3.1. This is because by the satistic theory of OLS estimator (Wooldridge , 2015), the sign and significance of $\hat{\beta}_{C,m}$ is sensitive to the so-called outlier housing units. While in the mid of the clustering process, the outliers to a given cluster C is also constantly changing along with the changing in the membership of C, which incurs the sequential dependence issue and makes it necesary to include the TBM procedure (algorithm TBM) into the iteration of clustering algorithm. On the other hand, by (11), the constraint (10) can be satisfied only if a sufficient large within-cluster co-variation trend exists between housing price and metro index. In the other words, the housing units grouped into one common cluster mustn't be quite similar to each other in terms of their non-spatial attribute, otherwise, the within-cluster co-variation trend cannot be present. This requirement causes the failure of the classical similarity-based clustering algorithm, such as the DBSCAN and K-means algorithms, in detecting clusters consistent with (10), for which we shall provide more numerical evidence in the following section.

4.2. Data

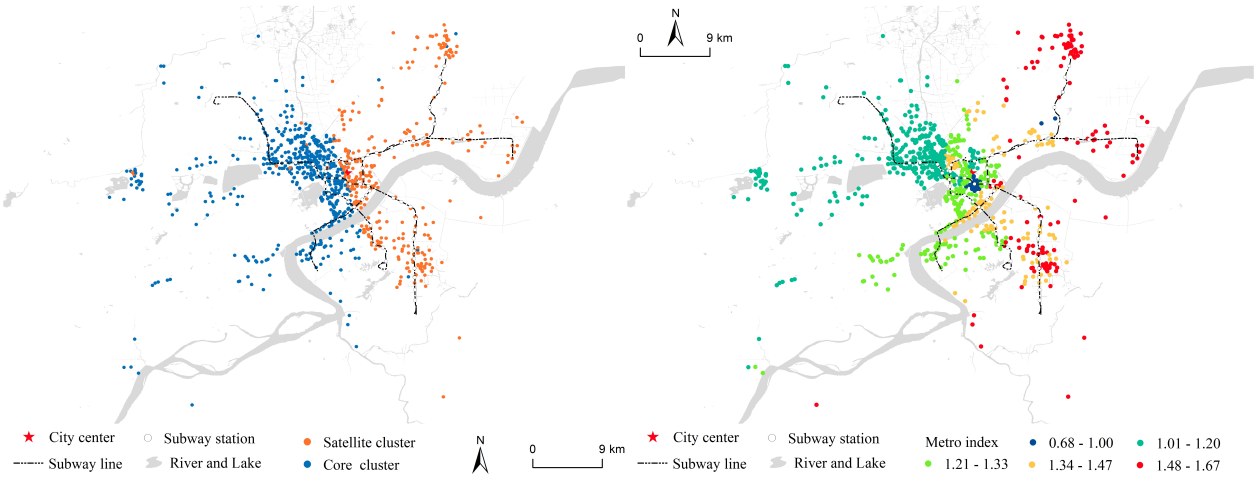
A one-week (Oct. 19 - 27, 2017) cross-sectional dataset of housing prices, housing types and subway accessibility is collected for Hangzhou, a large city in China. The data is collected from fang.com which is the largest and most famous online information platform

that provides detailed sales and transaction information of second-hand apartments in most cities of China (Li et al., 2017). The price on fang.com is updated timely to catch up with the market dynamics. Except for price, we also collect the building characteristics, such as area, building age and room structure, the neighborhood amenities, such as the green rate, volume rate, and school quality, and the location attributes, such distance to CBD, hospitals and commercial centers. All these attributes will be grouped into the control variables Z in the hedonic model (9). The latitude and longitude for each housing unit are looked up via Baidu geocoding API (http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-geocoding), based on which, the accessibility to subway of every housing unit is measured by the distance between the location of the unit and the nearest metro station. The metro index (8) is calculated for every housing unit to measure the degree of time saving by taking subway.

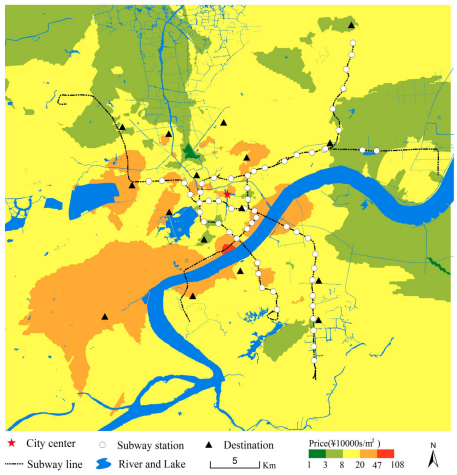
4.3. Clustering result

By applying algorithm TBM to the constrained clustering problem with constraint (10), we find that the optimal cluster number for Hangzhou is 2, which means heterogeneity of the housing premium induced by the improvement in transportation convenience does exist for Hangzhou. Meanwhile, the number of residual housing units is zero, which means the constraints (10) are perfectly satisfiable. The geographic range of the clustering results are plotted in Fig. 4a, in which the cluster covered by the blue dots includes the the urban core of Hangzhou, therefore, it is named as the "core" cluster; the other (covered by the orange dots) basically ranges over three satellite sub-cities, Xiasha, Xiaoshan and Linping, so it is called as the satellite cluster.

To facilitate the interpretation of the clustering result, we also plot the distribution of metro index and housing price in Fig. 4b and Fig. 4c, respectively. It is found from Fig. 4b that the geographic range of clusters is closely related with the spatial distribution of the metro index, which is plotted in Fig. 4b for comparison (Du et al., 2018; Li et al., 2017). Apparently, the range of the within-cluster variation of metro index plays an important role to determine the range of each cluster. In particular, the two clusters in Hangzhou almost coincide with the high-value region and low-value region in terms of the improvement degree of transportation convenience measured by metro index. In contrast to metro index, we find from Fig. 4c that the unit housing price provides almost zero information for identifying



a. Clusters for Hangzhou **b.** Metro index distribution



c. Housing price distribution

Figure 4: Clustering results and interpretation

In the application of Hangzhou housing market, the within-cluster homogeneity is defined via the positive transportation premium constraint (10). By this constraint, every cluster of the homogeneous housing units should share the identical co-variation trend between the price and the commuting efficiency brought by locating close to metro station. The co-variation-based clustering criteria differs significantly from the classical similarity-based clustering and cannot be implemented simply on the basis of the similarity distance between housing units attributes. This difference makes constrained clustering a indispensable tool for housing market segmentation.

the two clusters in Fig. 4a, because the core cluster and satellite cluster share almost the same variation range of the unit housing price. This observation suggests that the clustering algorithm with constraint (10) cannot only identify the submarkets consistent with the positive transportation premium criteria, but also works as a variable selection tool that can automatically identify the key variable to the satisfaction of the constraints.

From Fig. 4c, we also found that the within-cluster variation range for unit housing price is huge for both clusters in Fig. 4a, which suggests that if we take similarity-based clustering

algorithms (Ankerst et al., 1999; Birant and Kut, 2007; Guha et al., 1998; Zhang et al., 1996) to the data set, we may not be able to identify the two clusters in Fig. 4a. Because the within-cluster variation of data attributes are always relatively small for clusters identified by similarity-based clustering algorithms. As discussed in previous section, the positive within-cluster co-variation trend between housing price and metro index is naturally incompatible with the small within-cluster variation requirement, the similarity-based algorithms may not be able to find out appropriate clusters in terms of the constraint (10). To give a rigorous examination on this claim, we apply multiple similarity-based clustering algorithms to the Hangzhou's data set, including the density-based DBSCAN (Ester et al., 1996), ST-DBSCAN (Birant and Kut, 2007) and and the partition-based K-means algorithm. The clustering result is summarized in Table 1, where we report the total number K of clusters returned from each algorithm, the valid ratio (i.e. the ratio between the number of the clusters that have constraint (10) satisfied and K), the mean within-cluster variation range of unit housing price ($:= \frac{1}{K}\sum_{k=1}^{K}(\overline{P}_{C_k} - \underline{P}_{C_k})$ with $\overline{P}_{C_k}$ and $\underline{P}_{C_k}$ being the highest and lowest price within cluster C_k). The result shows that by all similarity-based clustering algorithm, most of the returned clusters fail to satisfy the positive premium constraints. Meanwhile the within-cluster price variation range resulting from the similarity-based algorithms is much smaller than the variation range (772,134 Chinese *yuan*) generated from our proposed algorithm. This observation supports our analysis on the incompatibility between the requirement on robust within-cluster co-variation trend and that on small within-cluster variation of housing attributes.

Table 1: Satisfaction to constraint (10) by alternative similarity-based clustering algorithms

Method	DBSCAN	ST-DBSCAN	K-Means			
Cluster number	611	610	2	3	4	5
Valid ratio	0.0769	0.077	0	0	0	0
Price variation range (unit: 10,000 *yuan*)	3.117	3.105	43.634	31.938	23.264	19.297

1. both spatial and non-spatial attribute are normalized by $(X - \mu_X)/\sigma_X$ where μ_X and σ_X are the mean and standard deviation of attribute X.

2. DBSCAN is implemented for a variety of tuning parameter values, where the radius *eps* varies within $(0.05, 0.1, \ldots, 1)$, the neighbor size *min_sample* varies from 2 to 10, the result is reported based on the parameters that minimize the size of the noisy set.

3. ST-DBSCAN is implemented for a variety of tuning parameter values, where both the spatial radius *eps*1 and non-spatial radius *eps*2 vary within $(0.05, 0.1, \ldots, 1)$, the neighbor size *min_sample* varies from 2 to 10, the result is reported based on the parameters that minimize the size of the noisy set.

Table 2: Regression Results By Clusters (Hangzhou)

Variable	Satellite Cluster Coef.	Core Cluster Coef.
distance to subway	**-0.033**	**-0.175*****
nearest metro less than 1km	**-0.017***	**-0.099****
nearest metro (1km,2km)	**0.023***	**-0.06***
metro index	**0.207*****	**0.185****
Adj. R^2	0.901	0.896
F-statistic	734.1***	997***
Obs.	1699	2431

$*$: 10% significant, $**$: 5% significant, $***$: 1% significant.

Table 2 reports the estimated coefficient for the accessibility variables to subway and metro index, which are calculated from estimating the hedonic model (9) within each sub-market/cluster. The regression coefficient of the accessibility measures are positive for the satellite cluster that has high-value metro index inside, and negative for the core cluster with low-value metro index. This observation implies that the housing price premium induced by being close to subway station is sensitive to the actual utility that the subway can bring to the residents, the more convenience induced by subway the higher premium it would contribute to the housing price and the vice verse. This finding agrees with the classical urban economic theory (Alonso et al., 1964; Mills, 1972; Muth, 1969), which highlights that the transportation convenience generates premium for housing units. However, in most empirical

studies, the accessibility to subway station is used as the only variable to characterize the subway premium, which neglects the potential spatial heterogeneity of subway stations in different regions in terms of their actual effect on improving local transportation efficiency.

Finally, we remark that the findings from Fig. 4b and Table 2 also provide deep insight into the urban structure of Hangzhou and its impact on the determinant of housing price. For instance, the core cluster in Fig. 4b reveals that, as a key feature of urban structure, the size of the core region of a city is exceptionally influential to the functionality of subway system. The city core of Hangzhou is tiny, it is centred at the Wulin Square and extends outward, reflecting as the region in Fig. 4b where the blue dots are gathered most densely. Also as shown in Fig. 4ba, due to the existence of the West Lake on the west, the Qiantang river on the south and east, and a mountain area in the northern-east, there exists natural boundaries for the expansion of the core region of Hangzhou. As a consequence, the urban core of Hangzhou is born to be small, it allows only a few metro stations to sit in it. Then, the accessibility to metro stations inside the core is poor, reducing the attractiveness of subway to local residents. On the other hand, the limitation of space restricts the speed advantage of subway in contrast to the ground traffics which further reduces the traffic demand to the underground subway system. This is because most of major destinations inside the urban core are not distant, which makes the ground traffic tools, such as bus, taxi and even bicycles, are not that inefficient compared to subway. Therefore, the size of city core is crucial to the comparative advantages of subway system and the demand to it, which further determines the difference of metro index and the heterogeneity in the metro premium of housing units between two submarkets in Hangzhou.

5. Conclusion

In this paper, the clustering tasks with implicit constraints are discussed. Compared with the traditional pairwise constraints and the size constraints, implicit constraints significantly increase the flexibility to include the user-specified homogeneity criteria into the the clustering task. The flexibility is critical to the applications in housing market studies, where it is lack of an universal criteria to define the homogeneity between two data units. Consequently, researchers have to adopt customized criteria in practice, which varies from case to case and rises up challenge to the algorithm design. Within implicit constraints, there are often complex inequalities involved, which makes the design of clustering algorithm generically more

complicated than that for the pairwise and/or size constraints. Particularly, the sequential dependence becomes a major issue when inserting data instances into clusters, which cause non-robust clustering result that is sensitive to the default order of data instances. To avoid the sequential sensitivity, a new algorithm is designed which uses the back-and-forth searching technique to minimize the impact of inserting order. The numerical experiment by the synthetic data demonstrates that the new algorithm outperforms the classical PCK-Means algorithm. As an application, we apply the new algorithm to a real dataset of housing price in Hangzhou. The clustering result implies useful information that is intuitively correct and therefore can be used as a verification of the classical economic theory.

There are also limitations in this paper: first, the back-and-forth searching for the appropriate clusters is computationally intensive, increasing the complexity of the algorithm. How to balance between the computation complexity and the order sensitivity is an important issue, especially for big data. This is not resolved in the current paper and deserve for future studies. Second, the back-and-forth searching can effectively handle the order sensitivity of the type "missing-inclusion". For the order sensitivity induced by the "including-exclusion", the current algorithm may not be a good remedy, new algorithm is needed. The recent development in deep learning indicates that the reinforcement learning techniques are efficient to solve many traditional combinatorial optimization problems(Fan et al., 2020) which provides us with some hints to the design of efficient solutions to the clustering problem with implicit constraints because constrained clustering is a special class of combinatorial optimization problems. Following this direction, some innovative work might be done in the future.

Compliance with Ethical Standards

Disclosure of potential conflicts of interest: N/A

Research involving Human Participants and/or Animals: N/A

Informed consent: N/A

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Alonso W. et al. Location and land use, 1964.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., Sander, J. OPTICS: Ordering points to identify the clustering structure. *In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pp. 4960, Philadelphia, USA, 1999. https://doi.org/10.1145/304181.304187

Bajic V. (1983). The effects of a new subway line on housing prices in metropolitan toronto. *Urban Studies*, 20(2):147–158. https://doi.org/10.1080/00420988320080291

Basu S., Banerjee A. and Mooney R.J. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 333–344. SIAM, 2004. https://doi.org/10.1137/1.9781611972740.31

Basu S., Davidson I. and Wagstaff K. *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press, 2008.

Birant, D., & Kut, A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 60(1): 208-221, 2007. https://doi.org/10.1016/j.datak.2006.01.013

Bonhomme S. and Manresa E. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015. https://doi.org/10.3982/ECTA11319

Bohman H. and Nilsson D. (2016). The impact of regional commuter trains on property values: Price segments and income. *Journal of Transport Geography*, 56:102–109. https://doi.org/10.1016/j.jtrangeo.2016.09.003

Bourassa S.C., Hoesli M., MacGregor B.D. et al. (1997). *Defining residential submarkets: evidence from Sydney and Melbourne*. HEC/Université de Genève.

Bourassa, Steven C and Hoesli, Martin and Peng, Vincent S (2003). *Do housing submarkets really matter?*. *Journal of Housing Economics*, 12:12–28. https://doi.org/10.1016/S1051-1377(03)00003-2

Can A. The measurement of neighborhood dynamics in urban house price. *Economic Geography*, 66(3):254–272, 1990. https://doi.org/10.2307/143400

Dao T., Duong K. and Vrain C. Constrained clustering by constraint programming. *Artificial Intelligence*, 244: 70-94, 2017. https://doi.org/10.1016/j.artint.2015.05.006

Davidson I. and Ravi S.S. Clustering with constraints: Feasibility issues and the K-Means algorithm. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 138–149. SIAM, 2005. https://doi.org/10.1137/1.9781611972757.13

Deng, M., Liu, Q., Cheng, T., & Shi, Y. An adaptive spatial clustering algorithm based on delaunay triangulation. *Computers, Environment and Urban Systems*, 35: 320332, 2011. https://doi.org/10.1016/j.compenvurbsys.2011.02.003

Dewees D.N. (1976). The effect of a subway on residential property values in toronto. *Journal of Urban Economics*, 3(4):357–369. https://doi.org/10.1016/0094-1190(76)90035-8

Diaz-Valenzuela I., Loia V., Martin-Bautista M.J., Senatore S. and Vila M.A. Automatic constraints generation for semisupervised clustering: experiences with documents classification. *Soft Computing*, 20(6):2329–2339, 2016. https://doi.org/10.1007/s00500-015-1643-3

Diego V.H., Paulina M. and Cesar F. Semi-supervised clustering algorithms for grouping scientific articles. *Procedia Computer Science*, 108:325–334, 2017. https://doi.org/10.1016/j.procs.2017.05.206

Du Q., Wu C., Ye X., Ren F. and Lin Y. Evaluating the effets of landscape on housing prices in urban China. *Tijdschrift voor economische en sociale geografie*, 109(4):525–541, 2018. https://doi.org/10.1111/tesg.12308

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with Noise. *In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 226231, Portland, OR, 1996.

Estivill-Castro, V., & Lee, I. Multi-level clustering and its visualization for exploratory spatial analysis. *GeoInformatica*, 6: 123152, 2002. https://doi.org/10.1023/A:1015279009755

Fan, C., Zeng, L., Sun, Y., and Liu, Y. Finding key players in complex networks through deep reinforcement learning. *Nature Machine Learning*, 2: 317–324, 2020. https://doi.org/10.1038/s42256-020-0177-2

Feng C.C., Wang F.L. and Gan L. Impact of rail transit on the residential property prices of submarkets: a case of the longgang line of shenzhen. *Progress in Geography*, 33(6): 765–772, 2014.

Fernando, B., Victor, L., & Marco, P. The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Computers & Geosciences*, 31(2): 155-163, 2005. https://doi.org/10.1016/j.cageo.2004.06.013

Fisher J., Christen P., Wang Q. and Rahm E. A clustering-based framework to control block sizes for entity resolution. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 279–288. ACM, 2015. https://doi.org/10.1145/2783258.2783396

Gabrielli L, Giuffrida S, Trovato M R. Gaps and overlaps of urban housing sub-market: hard clustering and fuzzy clustering approaches. In *Appraisal: from theory to practice*, pages. 203–219. Springer, Cham, 2017. http://dx.doi.org/10.1007/978-3-319-49676-4_15

Ganarski P., Dao T., Crmilleux B., Forestier G. and Lampert T. Constrained Clustering: Current and New Trends. *A Guided Tour of Artificial Intelligence Research*, pp. 447-484, 2020. https://doi.org/10.1007/978-3-030-06167-8_14

Goodman A.C. and Thibodeau T.G. Housing market segmentation. *Journal of housing economics*, 7(2):121–143, 1998. https://doi.org/10.1006/jhec.1998.0229

Galster George. Comparing demandside and supplyside housing policies: Submarket and spatial perspectives. *Housing Studies*, 12(4):561–577, 1997. https://doi.org/10.1080/02673039708720916

Goodman A.C. and Thibodeau T.G. Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, 12(3):181–201, 2003. https://doi.org/10.1016/S1051-1377(03)00031-7

Guha, S., Rastogi, R., Shim, K. CURE: An efficient clustering algorithm for large database. *In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 7384, USA, New York, 1998. https://doi.org/10.1145/276305.276312

Guo, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22 (7): 801823, 2008. https://doi.org/10.1080/13658810701674970

Guo K, Wang J, Shi G, et al. Cluster analysis on city real estate market of China: based on a new integrated method for time series clustering. *Procedia Computer Science*, 9: 1299-1305, 2012. https://doi.org/10.1016/j.procs.2012.04.142

Helbich M, Brunauer W, Hagenauer J, et al. Data-driven regionalization of housing markets. *Annals of the Association of American Geographers* 103(4): 871-889, 2013. https://doi.org/10.1080/00045608.2012.707587

Hepsen A, and Vatansever M. Using hierarchical clustering algorithms for turkish residential market. *International Journal of Economics and Finance*, 4(1): 138-150, 2012. https://doi.org/10.5539/ijef.v4n1p138

Hong Y. and Kwong S. Learning assignment order of instances for the constrained K-Means clustering algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):568–574, 2008. https://doi.org/10.1109/TSMCB.2008.2006641

Hu Q.W., Wang M., and Li Q.Q. (2014). Urban hotspot and commercial area exploration with check-in data. *Acta Geodaetica et Cartographica Sinica*, 43(3):314–321.

Hwang S, Thill J C. Delineating urban housing submarkets with fuzzy clustering. *Environment and Planning B: Planning and Design*, 36(5): 865-882, 2009. https://doi.org/10.1068/b34111t

Islam K.S. and Asami Y. Housing market segmentation: a review. *Review of Urban & Regional Development Studies*, 21(2-3):93–109, 2009. https://doi.org/10.1111/j.1467-940X.2009.00161.x

Kamw F., Shamal A.L., Zhao Y., Yang J., Ye X. and Chen W. Visually analyzing latent accessibility clusters of urban POIs. *EuroVA Workshop on Visual Analytics*, 2019. https://doi.org/10.2312/eurova.20191123

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. and Kumar, V. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017. https://doi.org/10.1109/TKDE.2017.2720168

Karypis, G., Han, E. H., & Kumar, V. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32: 6875, 1999. https://doi.org/10.1109/2.781637

Kiefer H. (2011). The house price determination process: Rational expectations with a spatial context. *Journal of Housing Economics*, 20(4):249–266. https://doi.org/10.1016/j.jhe.2011.08.002

Knight R. Baldassare M. and Swan S. (1979). Urban service and environmental stressor. *Environment and Behavior*, 11(4):435–450. https://doi.org/10.1177/0013916579114001

Le H.M., Eriksson A., Do T.T. and Milford M. A binary optimization approach for constrained K-Means clustering. In *Asian Conference on Computer Vision*, pages 383–398. Springer, 2018. https://doi.org/10.1007/978-3-030-20870-7_24

Li S., Ye X., Lee J., Gong J. and Qin C. Spatialtemporal analysis of housing prices in China: A big data perspective. *Applied Spatial Analysis and Policy*, 10(3): 421–433, 2017. https://doi.org/10.1007/s12061-016-9185-3

Li W., Feng C. and Zhao F. (2011). Influence of rail transit on nearby commodity housing prices: A case study of beijing subway line five. *Dili Xuebao/Acta Geographica Sinica*.

Liu, Q., Deng, M., Shi, Y., & Wang, J. A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Computers & Geosciences*, 46: 296-309, 2012. https://doi.org/10.1016/j.cageo.2011.12.017

Liu, Q., Liu, W., Tang, J., Deng, M., & Liu, Y. Two stage permutation tests for determining homogeneity within a spatial cluster. *International Journal of Geographical Information Science*, 33(9): 1718-1738, 2019. https://doi.org/10.1080/13658816.2019.1608998

McGreal S. and Paloma T. (2013). Implicit house prices: Variation over time and space in spain. *Urban Studies*, 50(10):2024–2043. https://doi.org/10.1177/0042098012471978

Miranda L., Viterbo-Filho J. and Bernardini F.C. RegK-Means: A clustering algorithm using spatial contiguity constraints for regionalization problems. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 31–36. IEEE, 2017. https://doi.org/10.1109/BRACIS.2017.70

Mills E.S. Studies in the structure of the urban economy, 1972.

Muth R.F. Cities and housing; the spatial pattern of urban residential land use, 1969.

National Bureau of Statistics. (2016). China City Statistical Yearbook. *Statistics Press*, Beijing: China.

Ottensmann J.R., Payton S., Man J. et al. (2008). Urban location and housing prices within a hedonic model. *Journal of Regional Analysis and Policy*, 38(1):19–35. http://dx.doi.org/10.22004/ag.econ.132338

Pels E., Debrezion G. and Rietveld P. (2010). The impact of rail transport on real estate prices. *Urban Studies*, 48(5):997–1015. https://doi.org/10.1177/0042098010371395

Plaut P.O. and Plaut S.E. (1998). Endogenous identification of multiple housing price centers in metropolitan areas. *Journal of Housing Economics*, 7(3):193–217. https://doi.org/10.1006/jhec.1998.0230

Pun L., Hui C.M., Chau C.K. and Law M.Y. (2007). Measuring the neighboring and environmental effects on residential property value: Using spatial weighting matrix. *Building and Environment*, 42(6):2333–2343. https://doi.org/10.1016/j.buildenv.2006.05.004

Randel R., Aloise D., Mladenović N. and Hansen P. On the k-medoids model for semi-supervised clustering. In *International Conference on Variable Neighborhood Search*, pages 13–27. Springer, 2018. https://doi.org/10.1007/978-3-030-15843-9_2

Rosen, S. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974. https://doi.org/10.1086/260169

Soaita, A.M., and Dewilde C. A critical-realist view of housing quality within the post-communist EU states: Progressing towards a middle-range explanation. *Housing, Theory and Society*, 36(1): 44-75, 2019. https://doi.org/10.1080/14036096.2017.1383934

Song Y., Zhang Y., Zheng S. and Zhong Y. (2016). The spillover effect of urban village removal on nearby home values in beijing. *Growth and Change*, 47(1):9–31. https://doi.org/10.1111/grow.12122

Tan, R., He, Q., Zhou, K., and Xie, P. The effect of new metro stations on local land use and housing prices: The case of Wuhan, China. *Journal of Transport Geography*, 79:102488, 2019. https://doi.org/10.1016/j.jtrangeo.2019.102488

Tang W., Yang Y., Zeng L. and Zhan Y. Size Constrained Clustering With MILP Formulation. *IEEE Access*, 2020. https://doi.org/10.1109/ACCESS.2019.2962191

Tang B. and Yiu C. (2010). Space and scale: A study of development intensity and housing price in hong kong. *Landscape and Urban Planning*, 96(3):172–182. https://doi.org/10.1016/j.landurbplan.2010.03.005

Tse R. (2002). Estimating neighbourhood effects in house prices: towards a new hedonic model approach. *Urban studies*, 39(7):1165–1180. https://doi.org/10.1080/00420980220135545

Wagstaff K., Cardie C., Rogers S., Schrödl S. et al. Constrained K-Means clustering with background knowledge. *Icml*, 1: 577–584, 2001.

Wang L. and Liu G. (2013). Spatial variation analysis of the housing price in multi-center city: A case study in chongqing city, china. In *Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on*, pages 450–453. IEEE. https://doi.org/10.1109/ICCIS.2013.125

Wang Y., Feng S., Deng Z., and Cheng S. (2016). Transit premium and rent segmentation: A spatial quantile hedonic analysis of shanghai metro. *Transport Policy*, 51:61–69. https://doi.org/10.1016/j.tranpol.2016.04.016

Wang X. (2017). Community attributes and the subway capitalization effect.

Wooldridge, Jeffrey M. Introductory econometrics: A modern approach. Cengage learning, 2015.

Wu C. and Sharma R. Housing submarket classification: The role of spatial contiguity. *Applied Geography*, 32(2):746–756, 2012. https://doi.org/10.1016/j.apgeog.2011.08.011

Wu C., Ye X., Ren F., Du Q. and Luo P. Spatial effects of accessibility to parks on housing price in Shenzhen, China. *Habitat International*, 63: 45–54, 2017. https://doi.org/10.1016/j.habitatint.2017.03.010

Wu C., Ye X., Ren F. and Du Q. Modified data-driven framework for housing market segmentation. *Journal of Urban Planning and Development*, 114(4): 04018036, 2018. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000473

Wu C., Ye X., Ren F., Wan Y., Ning P. and Du Q. Spatial and social media data analytics of housing prices in Shenzhen, China. *PloS One*, 11(10): e0164553, 2016. https://doi.org/10.1371/journal.pone.0164553

Wu, Y., Wei, Y.D., and Li, H. Analyzing spatial heterogeneity of housing prices using large datasets. *Applied Spatial Analysis and Policy*, 13(1): 223-256, 2020. https://link.springer.com/article/10.1007/s12061-019-09301-x

Yao J. and Fotheringham A.S. (2016). Local spatiotemporal modeling of house prices: a mixed model approach. *The Professional Geographer*, 68(2):189–201. https://doi.org/10.1080/00330124.2015.1033671

Ye X., Jiang Y., and Wang Z. (2007). Impact area of shanghai rail transit line 1 on development benefits. *Urban Mass Transit*, 10(2):28–31.

Yu, D., Yin, J., and Ye, F., 1 Novel methods to demarcate urban house submarket - cluster analysis with spatially varying relationships between house value and attributes. *IET International Conference on Smart and Sustainable City (ICSSC 2011)*, 2011. https://doi.org/10.1049/cp.2011.0288

Ye X., She B. and Benya S. Exploring regionalization in the network urban space. *Journal of Geovisualization and Spatial Analysis*, 2(1): 4, 2018. https://doi.org/10.1007/s41651-018-0013-y

Yu D., Wei Y., and Wu C. (2007). Modeling spatial dimensions of housing prices in milwaukee, wi. *Environment and Planning B: Planning and Design*, 34(6):1085–1102. https://doi.org/10.1068/b32119

Yu Z., Luo P., You J., Wong H., Leung H., Wu S., Zhang J. and Han G. Incremental semi-supervised clustering ensemble for high dimensional data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):701–714, 2015. https://doi.org/10.1109/TKDE.2015.2499200

Zhang D., Zhang X., Zheng Y., Ye X., Li S. and Dai Q. Detecting intra-urban housing market spillover through a spatial markov chain model. *ISPRS International Journal of Geo-information*, 9(1): 56, 2020. https://doi.org/10.3390/ijgi9010056

Zhang, T., Ramakrishnan, R., Livny, M. BIRCH: An efficient data clustering method for very large databases. *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 103114, Canada, Montreal, 1996. https://doi.org/10.1145/235968.233324

Zhang X., Zheng Y., Sun L. and Dai Q. Urban structure, subway system and housing price: Evidence from beijing and hangzhou, china. *Sustainability*, 11(3):669, 2019. https://doi.org/10.3390/su11030669

Zhu S., Wang D. and Li T. Data clustering with size constraints. *Knowledge-Based Systems*, 23(8):883–889, 2010. https://doi.org/10.1016/j.knosys.2010.06.003