A Clustering Method for Analysis of Data Subject to Pre-defined Classifications

Yang Liu[1]

Abstract In this paper, we present a methodology to perform clustering and grouping analysis for dataset with classification constraints or definitions. The discussion is demonstrated with a full example based on real data. We start with the observed difference in the CIA and UN subregional definition of European countries, and consider what the impact is from a subregional house price ratio perspective. As documented in this report, we find that the presented approach useful for clustering analysis of the pre-identified subgroups to address subgroup based clustering problems.

Keywords: *Data Classification, Grouping Analysis, Clustering*

Contents

1 Motivation

Clustering is an important topic in data analysis. While there are a lot of discussions around the methodologies of classification and pattern recognition, most discussions are based on single record level and assume that the difference between records are completely embedded in the features of the underlying data.

Therefore it is challenging when classification problem involves constraints or predefined data properties in the underlying data.

In this paper we propose an approach to perform classification analysis for data with constraints

[1] **Yang Liu** is a quantitative specialist at an international bank. Yang holds a doctorate in quantitative finance from Cass Business School, City University of London. He has published a number of papers on quantitative methods in risk and finance and served as reviewer for journals in the field.
The opinions expressed in this paper are those of the author only.
E-mail: yang.liu.q-fin@outlook.com

or pre-defined properties in place. To help demonstrate the approach, we provide a full analytical example using real data along side the methodology discussion.

We detail the data, target and challenge in the following sub-sections.

Data

Here we download and use the 2018 Q2 IMF house price-to-income and price-to-rent data for research and academic purpose only. Any conclusions and findings presented in this paper using the dataset does not reflect or contradict any conclusion presented or found in other publications.

Source of housing price ratio data we use in this paper is: IMF Global Housing Watch [1].

Note that the IMF analysis quote the source as Organisation for Economic Co-operation and Development, while we fully acknowledge the analysis and reports published by IMF, here we only make direct reference to the IMF web location where the data is downloaded.

We then filter the data and only keep the European countries to meet our focus of analysis.

As this data is at country level, we adopt and mapped these countries according to the European country classification rules of both the CIA and United Nations.

The CIA classification can be found in the World factbook: CIA World Factbook [2]. The link to the United Nations geographical classification definition is: United Nations Geoscheme [3].

The full final dataset used to demonstrate the methodology is as shown in Table 1.

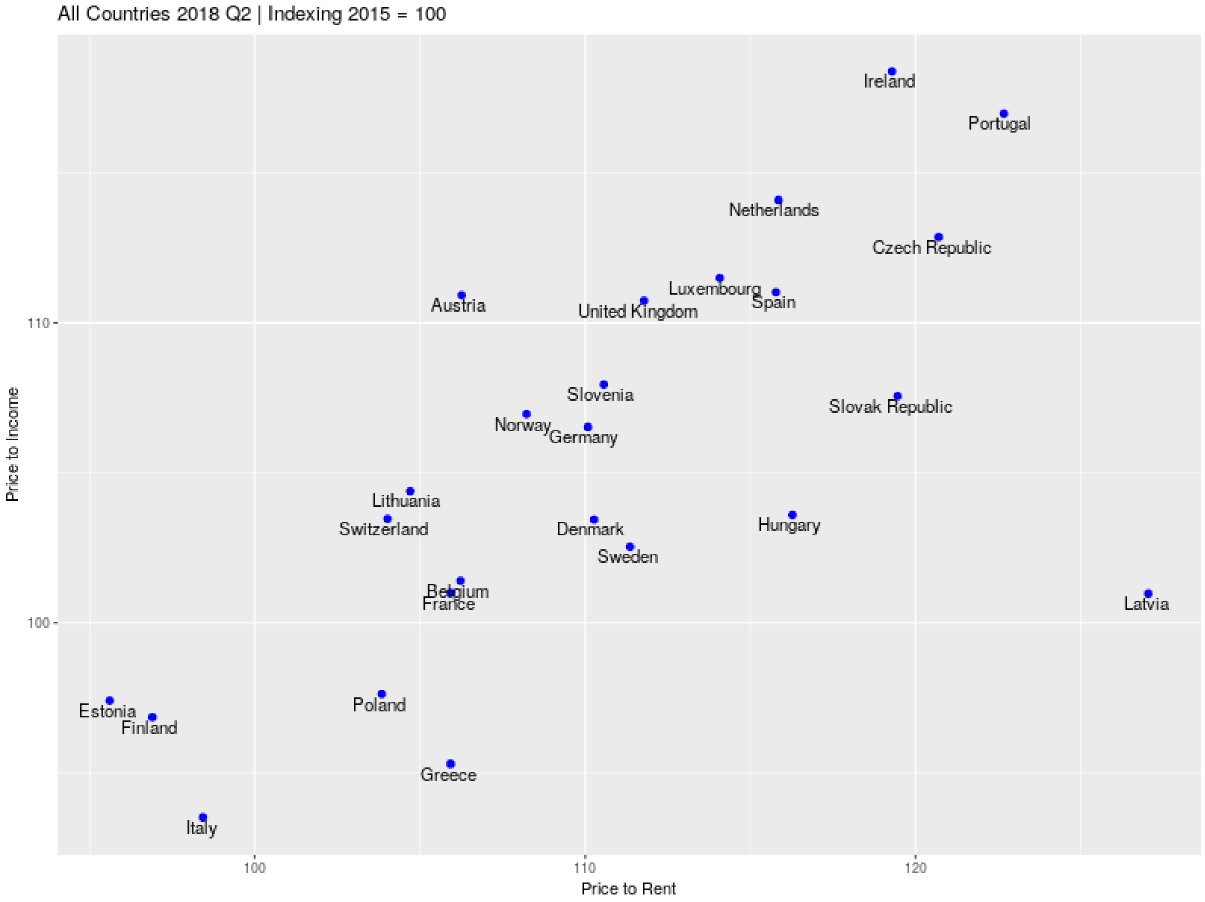
Table 1: 2018 Q2 Indexed (2015=100) IMF House Price-to-Rent and Price-to-Income Ratio with CIA and UN Geographical Region Classification for European Countries

Country	CIARegion	UNRegion	price_to_rent	price_to_income
Austria	CE	WE	106.26	110.91
Belgium	WE	WE	106.22	101.40
Czech Republic	CE	EE	120.70	112.85
Denmark	NE	NE	110.27	103.44
Estonia	EE	NE	95.61	97.41
Finland	NE	NE	96.90	96.86
France	WE	WE	105.94	100.99
Germany	CE	WE	110.08	106.52
Greece	SE	SE	105.93	95.30
Hungary	CE	EE	116.28	103.59
Ireland	WE	NE	119.28	118.36
Italy	SE	SE	98.44	93.52
Latvia	EE	NE	127.04	100.97
Lithuania	EE	NE	104.70	104.39
Luxembourg	WE	WE	114.07	111.49
Netherlands	WE	WE	115.85	114.08
Norway	NE	NE	108.22	106.95
Poland	CE	EE	103.85	97.62
Portugal	SE	SE	122.67	116.96
Slovak Republic	CE	EE	119.46	107.55
Slovenia	CE	SE	110.57	107.94
Spain	SE	SE	115.77	111.01
Sweden	NE	NE	111.37	102.54
Switzerland	CE	WE	104.03	103.46
United Kingdom	WE	NE	111.78	110.73

Note that the price ratios are indexed by basing the 2015 value as 100, we keep only 2 decimal places for display purposes in the table, original value is kept and used in calculations in the rest of this paper.

Figure 1 graphically display the raw data.

Figure 1: IMR House Price-to-Income and Price-to-Rent Index 2018 Q2 (2015=100)



Target

We acknowledge the rational respectively behind the CIA and UN classification of the countries. Our analytical target is to demonstrate the the analytical approach by establish a grouping and measuring system using these data and compare the results from a house price ratio perspective. Results and conclusion in this paper does not contradict the published CIA and UN classifications.

As shown in Table 1, we see that there are significant differences between the CIA and UN defined subregions of European countries. The UN consider four subregions: Eastern, Western, Northern and Southern, the CIA has a fifth subregion: Central Europe.

Aside from the total number of subregions, we see that there are differences in the classification of certain countries as well. For example, the UK is considered as a Western European country by CIA while it is a Northern European country to the UN, same for Ireland.

In the meantime, CIA's Eastern European countries Latvia and Estonia are considered as Northern European countries according to the UN. Additionally, Western European country Germany for the UN is classified as a Central European country by the CIA. More differences like these can be found in Table 1.

Now if we consider the house price ratios on a subregional basis in Europe, how does the above differences affect the result? Note that in this example the key difference is two sets of definitions on exactly the same underlying data.

The goal of this example is:

- First to demonstrate our approach to group and measure the cluster of data under pre-defined conditions, in this case the two different classification definitions.
- Second, we look into further clustering of pre-defined data groups, for example, if we separate the data into two groups only by further grouping the CIA or UN defined regions.
- Thirdly and specific to this example, we want to compare the two set of definitions from a housing price perspective using the IMF reported index data.

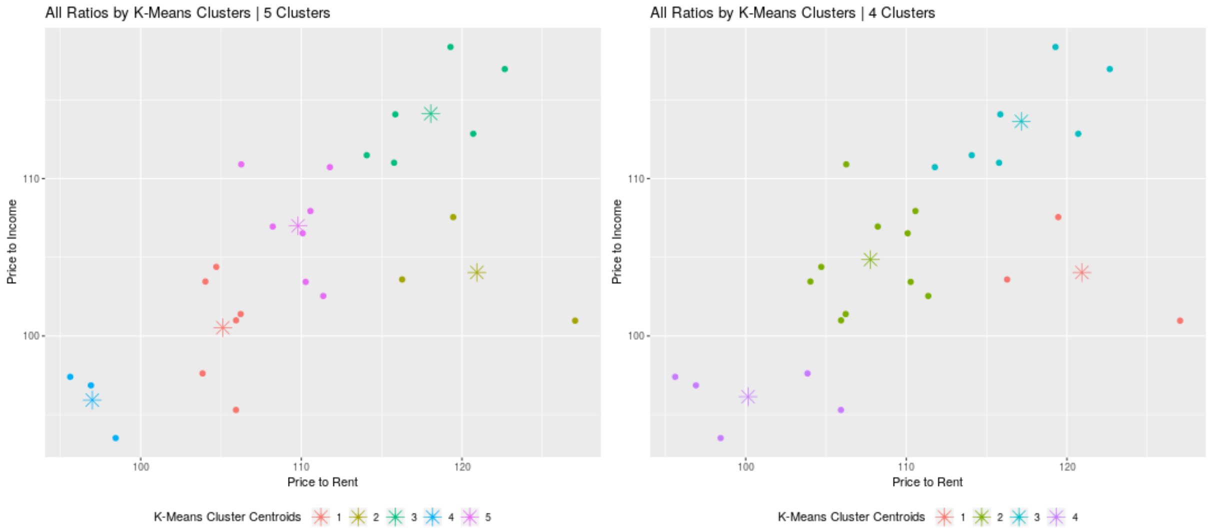
Challenge

Classic clustering classification method such as the K-means clustering work less well for the purpose of this example as the output clusters are strongly subject to constraints. As a result, we find the following:

- The suggested clusters does not necessarily complain with the pre-defined regions.
- For data record values that are numerically close, clustering results are subject to random seeds and processes.
- In case of further clustering or splitting of grouped data like grouping 5 geographic subregions into 2, is either not possible or violating the initial definition such as region classifications.

Figure 2 demonstrate the K-Means clustering of 5-cluster and 4-cluster classification respectively.

Figure 2: K-Means Clusters



2 Geometric Representation of Data

We adopt condition according to the pre-defined property or constraint, and geometrically represent the subset data as convex polygons. On a 2D plain, the XY-axis are the 2 numerical ratios that is considered in the analysis. Our approach works well in higher dimensions and adopt orthogonal transformation or dimension reduction methodologies, but we focus on the 2-dimensional example for simplicity and illustration purpose in this paper.

Tables 3a and 3b define the 2D polygons according to the CIA and UN definitions.

Table 2: Vertices for 2D Graphics Plot

Country	CIARegion
Czech Republic	CE
Slovak Republic	CE
Hungary	CE
Austria	CE
Switzerland	CE
Poland	CE
Latvia	EE
Lithuania	EE
Estonia	EE
Sweden	NE
Norway	NE
Finland	NE
Portugal	SE
Spain	SE
Greece	SE
Italy	SE
Ireland	WE
Luxembourg	WE
United Kingdom	WE
France	WE

(a) CIA Region Vertices

Country	UNRegion
Czech Republic	EE
Slovak Republic	EE
Hungary	EE
Poland	EE
Latvia	NE
Ireland	NE
Finland	NE
Estonia	NE
Portugal	SE
Slovenia	SE
Greece	SE
Italy	SE
Netherlands	WE
Luxembourg	WE
Austria	WE
France	WE
Switzerland	WE

(b) UN Region Vertices

Figure 3 and 4 illustrate the 2D polygons of the CIA and UN classification of European countries.

Figure 3: 2D polygons: CIA classification

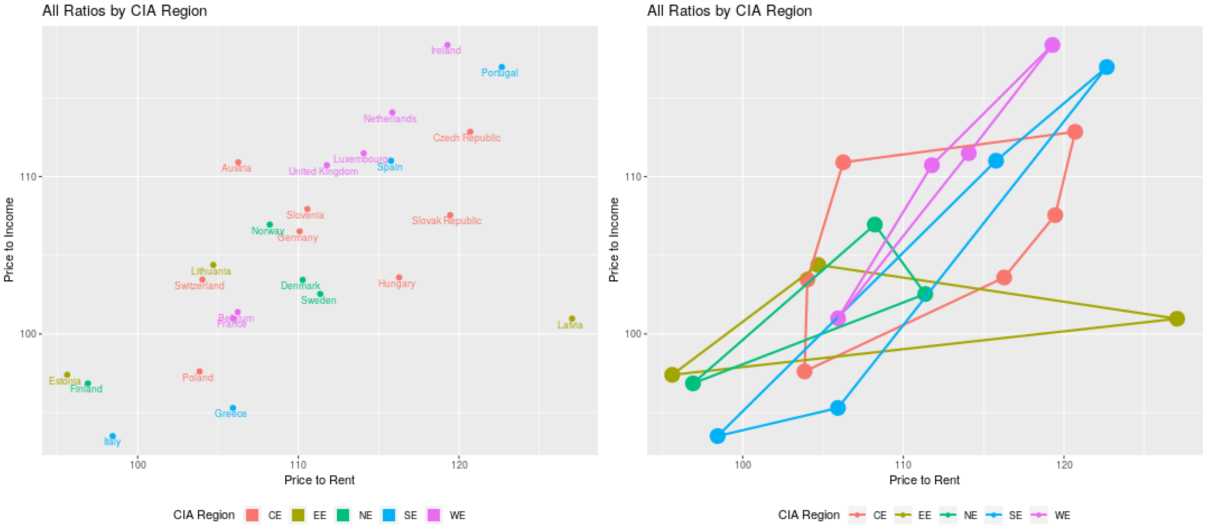
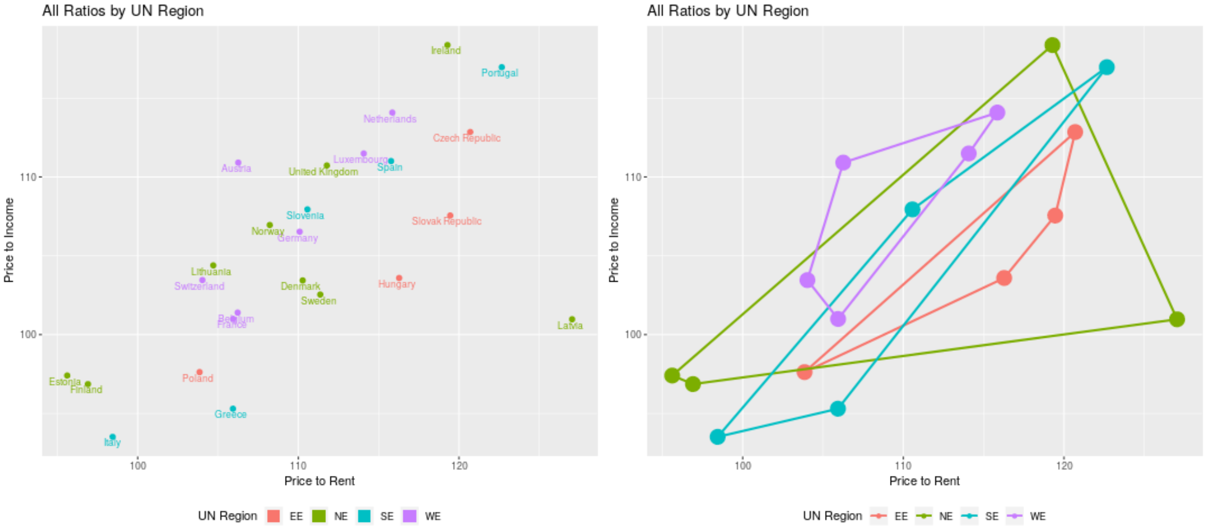


Figure 4: 2D polygons: UN classification



3 Relative Location of Geometric Representations

2D Geometric Representation

Given coordinates and XY-axis measurements, it is easy to assess the relative position of the convex polygons. In particular, we are interested in whether the polygons intersect each other, and if so, the impact of such intersections.

Tables 3 and 4 below summarize the relative locations between clusters within the CIA or UN classifications. The text note 'Cross' in the tables means the polygons intersect, and 'Out' indicates that the polygons are apart from each other on the plain.

Table 3: Relative Location on 2D Plain, CIA Regions

CIA Region	EE	NE	SE	WE
CE	Cross	Cross	Cross	Cross
EE		Cross	Cross	Cross
NE			Cross	Cross
SE				Cross

Table 4: Relative Location on 2D Plain, UN Regions

UN Region	NE	SE	WE
EE	Cross	Cross	Out
NE		Cross	Cross
SE			Cross

Definition: 2D Properties

It is known that the centroid of a convex polygon is within the edges of the polygon, also, the segment connecting the centroids of two non-intersecting polygons cross each polygon once and only once.

Denote the coordinates of data points using x and y, the centroid of a non-self-intersecting closed polygon defined by n vertices $(x_0,\ y_0)$, $(x_1,\ y_1)$, ..., $(x_{n-1},\ y_{n-1})$ is the denoted by $(C_x,\ C_y)$, where

$$C_{\mathrm{x}} = \frac{1}{6A}\sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i\ y_{i+1} - x_{i+1}\ y_i), \quad \text{and} \tag{1}$$

$$C_{\mathrm{y}} = \frac{1}{6A}\sum_{i=0}^{n-1}(y_i + y_{i+1})(x_i\ y_{i+1} - x_{i+1}\ y_i) \tag{2}$$

and calculate the polygon's signed area A as:

$$A = \frac{1}{2}\sum_{i=0}^{n-1}(x_i\ y_{i+1} - x_{i+1}\ y_i) \tag{3}$$

In these formulas, the vertices are ordered by their occurrence along the polygon's perimeter; furthermore, vertex (x_n, y_n) is the same as vertex (x_0, y_0), meaning the last vertex count loop back to $i = 0$ to close the polygon. The numerical result from these formulas can be both positive or negative based on the how the points are ordered, the sign has no other meaning aside from ordering.

4 General and Relative Distances

Relative Distance

We define the distance between any two polygons as following:

- if intersect, denote the distance as negative.

  The relative numerical value of the distance of polygon B to polygon A is the intersect area weighted by the total area of polygon A; the relative distance of A to B is the area of intersect weighted by area of B.

  This applies to the case where one polygon is completely positioned within the footprint of another, in which case the 'distance' of the small polygon to the big one is a value between 0 and -1, while the 'distance' of the big polygon to the small one is exactly -1.

- if non-intersect, the relative distance is positive.

  We define the relative distance of polygon B to A by the ratio of: the absolute length of the segment between the two intersect points on the edges, over the absolute length from centroid of A to intersect point on edge of polygon B.

Tables 5 and 6 below summarize the relative distances calculated for clusters within the CIA and UN classifications.

Table 5: Relative Distance between CIA Regions

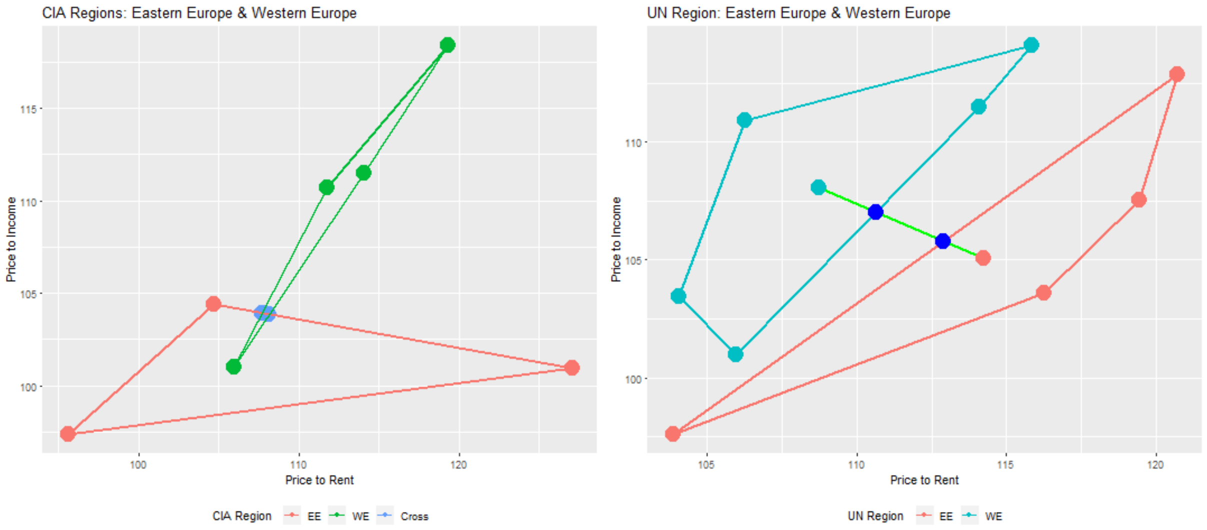
CIA Region	CE	EE	NE	SE	WE
CE	0.0000	-0.2399	-0.1896	-0.3117	-0.0678
EE	-0.3839	0.0000	-0.3402	-0.2364	-0.0079
NE	-0.6953	-0.7797	0.0000	-0.2391	-0.0469
SE	-0.6229	-0.2952	-0.1302	0.0000	-0.0001
WE	-0.6838	-0.0497	-0.1288	-0.0006	0.0000

Table 6: Relative Distance between UN Regions

UN Region	EE	NE	SE	WE
EE	0.0000	-0.9994	-0.3330	0.5467
NE	-0.1686	0.0000	-0.2572	-0.1334
SE	-0.1695	-0.7759	0.0000	-0.0698
WE	0.6320	-0.6839	-0.1185	0.0000

Figure 5 below showcase the above descriptions.

Figure 5: Relative Distances: Intersecting vs. Non-Intersecting



The Tables 5 and 6 are structured by row, where each row is the base subregion, the table is read as 'column-to-row', e.g. for the CIA subregions, the relative distance of 'EE' to 'CE' is -0.2399, while the relative distance of 'CE' to 'WE' is -0.6838.

Similarly in Table 6, we see that under UN definition, the region'EE' is almost entirely wrapped by 'NE', hence the 'NE' is of relative importance -0.9994 to 'EE'.

As mentioned before, the negative sign here only indicate that the 2D polygons intersect with each other.

General Distance

With the relative distances calculated, 'relative importance' analysis can be done easily using these measures. Meanwhile, the general distance on a 2D plan can be obtained from averaging the relative distances:

$$D_{A,B} = \frac{\widetilde{D}_{A\leftarrow B} + \widetilde{D}_{B\leftarrow A}}{2}$$

Here $D_{A,B}$ is the general distance between polygons A and B, $\widetilde{D}_{A\leftarrow B}$ is the relative distance of B to A. Note that this distance comes in vector form in higher dimensions. The above equation represents a 2D distance calculation which is essentially a matrix operation that can be useful in higher dimension scenarios.

Tables 7 and 8 below summarize the total distance between subregions within the CIA and UN classifications.

Table 7: Distance between CIA Regions

CIA Region	CE	EE	NE	SE	WE
CE	0.0000	-0.3119	-0.4424	-0.4673	-0.3758
EE	-0.3119	0.0000	-0.5599	-0.2658	-0.0288
NE	-0.4424	-0.5599	0.0000	-0.1847	-0.0878
SE	-0.4673	-0.2658	-0.1847	0.0000	-0.0003
WE	-0.3758	-0.0288	-0.0878	-0.0003	0.0000

Table 8: Distance between UN Regions

UN Region	EE	NE	SE	WE
EE	0.0000	-0.5840	-0.2512	0.5894
NE	-0.5840	0.0000	-0.5165	-0.4087
SE	-0.2512	-0.5165	0.0000	-0.0941
WE	0.5894	-0.4087	-0.0941	0.0000

Clustering with Constraints

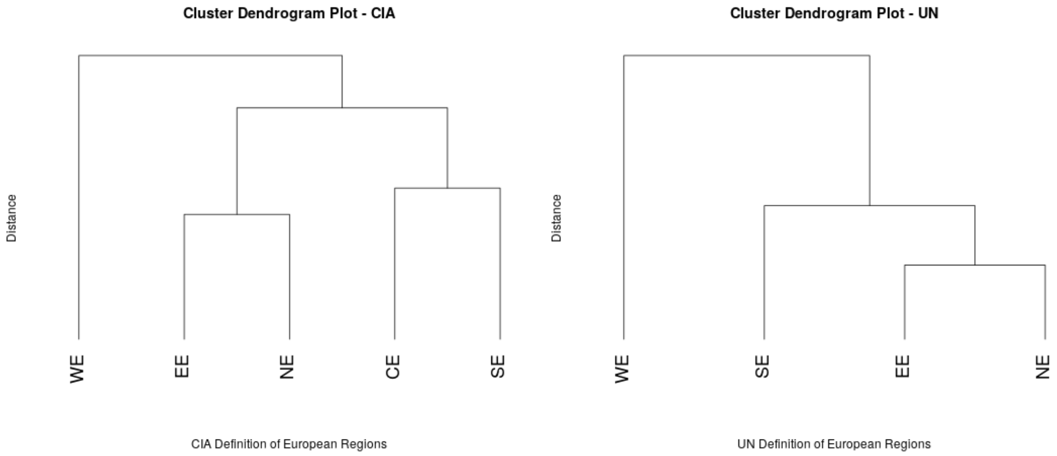
When geographical or industrial definitions are introduced, it is often the perspective from which the answer is expected. Given the subregion definitions in our example, it is intuitive to ask the following questions:

- if we were to categories Europe into 2 subregions only, which of the existing subregions shall we group together under CIA and UN definitions, respectively?
- given CIA has 5 and UN has 4 subregions defined, which subregions can we group together to reduce the number of subregions by 1?

The linkage dendrogram plots can be obtained easily for clustering analysis with constraints.

Figure 6 demonstrate the links and path of clustering results using the general distance between the pre-defined subregions under CIA and UN definitions respectively.

Figure 6: Clustering Dendrogram Plots: CIA and UN Definitions



Conclusion of the Example

From a house price ratio perspective, visual comparison using Figures 4 and 6 show that, under the UN definition, the Western Europe is completely apart from Eastern Europe and more distinguishable from the rest of subregions. Whilst under the CIA definition the Western Europe has least intersection with the other subregions, it is less clear comparing to the UN scenario.

In the meantime, we see that the inclusion of countries such as UK, Ireland and Latvia in the UN geographical definition has changed the coverage and data diversification of Northern Europe significantly, it is therefore making this subregion less distinguishable from the others as it is now intersecting with ever other UN subregions.

Revisit the simple questions mentioned in previous section, we are now able to address them with results obtained so far.

- if we were to categories Europe into 2 subregions only, which of the existing subregions shall we group together under CIA and UN definitions, respectively?

  A: Inspecting Figure 6 top-down from the ‘roots’ down to the ‘leaves’, the analysis suggest a ‘west and rest’ classification following both CIA and UN definitions.

- Given CIA has 5 and UN has 4 subregions defined, which subregions can we group together to reduce the number of subregions by 1?

  A: Look at the Figure 6 bottom-up from the ‘leaves’ upwards, the analysis suggest combining Eastern and Northern Europe as a first step for both the CIA and UN definition.

  Note that one step further under the CIA definition would require combination of Central Europe with the Southern Europe in order to reduce the total number of subregions to 3.

Other Use Cases

Aside from the example , the potential use case of this methodology also include:

- Impact analysis of single record changes between different definition settings. Back to our example, what will happen in the subregional housing price ratios if Latvia was classified as an Eastern European country under the UN definition instead of Northern.
- Constrained clustering and aggregation by pre-classified properties. Take a corporate lending portfolio for example, what is the regional classification of portfolio defined using country of incorporation vs. country of operation? Another example is when the analysis require grouping data with pre-defined industry sectors by identifying whether mining industry customers behave more like customers in the energy sector or fishing sector.
- Outlier testing. It is obvious that the geometrical representation of data changes with the inclusion or exclusion of data points as vertices. Therefore, the presented approach can be used for outlier identification and adjustments while considering all other data records observed in the same classification.

5 Conclusions

In this paper, we present an approach to perform clustering analysis when the target is subject to pre-defined subcategories in the underlying dataset. As an illustrative example, we perform full analysis on real data to address real world questions.

With the proposed methodology, one is able to quantify and measure relative and general distances between pre-defined subcategories within the dataset, hence quantitative clustering analysis conditional on the pre-defined subcategories are made possible even if the subcategories are not defined from a perspective that is related with the underlying data.

References

1 International Monetary Fund *IMF Global Housing Watch: House Price-to-Income Ratio, House Price-to-Rent Ratio.* URL https://www.imf.org/external/research/housing/.

2 Central Intelligence Agence *Wikipedia Page of European Subregion Map: CIA World Factbook.* URL https://en.wikipedia.org/wiki/Western_Europe#/media/File:Europe_subregion_map_world_factbook.svg.

3 United Nations Geoscheme *Wikipedia Page of European Subregion Map: United Nations geoscheme for Europe.* URL https://en.wikipedia.org/wiki/United_Nations_geoscheme_for_Europe#/media/File:Europe_subregion_map_UN_geoscheme.svg.