Matching and Clustering in Square Contingency Tables: Who Matches with Whom in the Spanish Labour Market?

Pablo Álvarez de Toledo Saavedra
pabloat@us.es
Departamento de Organización Industrial y Gestión de Empresas I.
Universidad de Sevilla.

Fernando Núñez Hernández
fnunezh@us.es
Departamento de Organización Industrial y Gestión de Empresas I.
Universidad de Sevilla.

Carlos Usabiaga Ibáñez
cusaiba@upo.es
Departamento de Economía, Métodos Cuantitativos e Historia Económica.
Universidad Pablo de Olavide.

Subject area:

10. Mercado de Trabajo y Territorio.

Abstract:

In this study we develop previous contributions on segmentation in the Spanish labour market (Álvarez de Toledo, Núñez and Usabiaga, 2013, 2014, 2017) inserting their contents in the more general framework of contingency tables and the dimension problem generated by the combination of attributes of multiple variables. After presenting the characteristics of the contingency table which is our object of study, we develop the concepts of propensity to match between each row category and each column category and of similarity between two row (or column) categories. Using a Cobb-Douglas functional form we estimate the global propensity to match between each row category and each column category as a multiplicative function of the partial propensities to match between attributes of rows and attributes of columns (one by one); similarly, we can decompose the measure of similarity. Following this approach, we disentangle the mix of the effects of the multiple interactions between the different attributes. We also apply biclustering procedures to our contingency table using our measure of similarity.
The methodology is tested through an application to matching data of the Spanish labour market (Continuous Sample of Working Lives, Muestra Continua de Vidas Laborales, MCVL). To define the labour categories, we use the same variables for workers and jobs; i.e., province and occupation during the search and matching process. We represent the original contingency table as a “matching map” of biclusters by means of a clustering process and show, for each obtained bicluster, the corresponding propensity to match between workers and jobs. In general, we observe a clear relationship between the clusters obtained for combined province-occupation categories and those obtained for provinces and occupations separately. Our analysis shows that worker mobility (geographical or occupational) and the availability of enough useful information are important requirements for effective labour matching.

**Keywords:**
Labour Matching, Contingency Tables, Segmentation, Cluster Analysis, Continuous Sample of Working Lives, Spanish Labour Market.

**JEL codes:**
J61, J62, J63, J64, C38, C55.
1. Introduction

Let us suppose that, in a population isolated from the rest of the world, we know the annual number of marriages between men and women, between fat men and thin women and between short men and tall women. What will be the annual number of marriages between fat and short men and thin and tall women? Obviously, this is a question of dependence or independence between variables.

This kind of questions is present in this paper which is a substantial development of our previous papers, particularly with regard to two points. First, we insert their contents in the more general framework of contingency tables in multivariate statistics. Second, we deal with the ‘dimensions problem’ generated by the combination of the multiple characteristics that define each row and column categories (workers and jobs categories in the application of section 3) and make their number soar. We propose a factor approach to disentangling the mix of the effects of the multiple interactions between the different variables. Let us examine both issues more closely.

N-way contingency tables (also known as cross-classification or cross-tabulation tables) are a statistical tool for summarising and displaying the relation between $n$ categorical variables$^1$. The cells of the table represent all the possible combinations of the different categories of each variable, and each cell contains the corresponding frequency count. In the case of two-way contingency tables, we have a rectangular table or matrix, with rows and columns corresponding respectively to the different categories of two variables. A two-way contingency table in which there is a one-to-one correspondence between the categories of the row and column variables is a square contingency table. Thus, $r = c$ ($r$ is the number of rows and $c$ is the number of columns in the table). Square contingency tables are common in many fields: switching matrices in market research that show purchasing or consumption of brands on two separate occasions; ‘treatments’ comparison or ‘before-after’ comparisons; social mobility; opinion surveys; rater agreement, etc. In many cases the frequencies on the ‘main diagonal’ (the cells for which the row category is the same as the column category) are relatively large. This indicates, respectively, that most buyers did not change their brand choice, small change before-after, low mobility, little change of opinion, raters rather agree, etc.

---

$^1$ Contingency tables have been treated extensively in academic literature. See, for example, Bishop et al. (1975), Fienberg (1980) and Agresti (2013). For a historical perspective, see Stigler (2002), Fienberg and Rinaldo (2007) and Agresti (2007, chapter 11).
In previous papers –see Álvarez de Toledo et al. (2014, 2017)–, as well as in our application of section 3, data on observed matches (frequency counts) in the Spanish labour market for different categories of workers and jobs (categorical variables) are used. We consider the same categories for workers and jobs (square contingency tables), which allows us to make a simpler analysis of mobility and to obtain clearer results in the clustering process. Additionally, the analysis we develop here could be extended to non-square tables.

More generally, contingency tables can be related also to data on observed matches in one-to-one two-sided matching markets (‘who matches with whom’), which are cross-classified by some characteristics of the parties in the match\(^2\). If the characteristics considered in the two sides and their categories are the same, we will have a square contingency table. Again, in these type of markets, large frequencies in the main diagonal are indicative of homogamy (in the marriage market, tendency to marry someone similar to oneself), low mobility (geographical and occupational in the labour market), etc.

However, in general, the relation between contingency tables and two-sided matching literatures is small. Contingency tables focus directly on data analysis, whereas in two-sided matching, data analysis is generally founded on previous theoretical models, and contingency tables do not appear or appear marginally –see, for example, Graham (2011), Agarwal (2015) and Chiappori and Salanié (2016). On a few occasions we can find a greater relation as, for example, in Hsieh (2012) and, to a lesser extent, in Arenas (2014) and Goñi (2015). Two-sided matching centres on the question of who matches with whom in a one-to-one correspondence between each of the agents in the two sides of the market, which is denoted as matching outcome, observed matches, ‘who matches with whom’, etc. To get a contingency table requires additionally to cross-classify the agents in categories by some characteristics, and to compute the frequencies for each combination of categories.

The notion of *propensity to match* between each row (worker) category and each column (job) category in two-way contingency tables plays a central part in our paper. It can be related to the notion of departure from independence in contingency tables (difference between the

\(^2\) The vast two sided matching literature begins with the seminal work of Gale and Shapley (1962) and Becker (1973, 1991). Since then, this field has generated many theoretical and empirical papers, laboratory experiments, textbooks, surveys and a great variety of market design applications. The survey of Roth and Sotomayor (1990) covers the literature on two-sided matching markets until 1990. More recent surveys are Abdulkadiroglu and Sönmez (2013), and, specifically on the marriage market, Browning et al. (2014). Chiappori and Salanié (2016) survey the empirical/econometric work. From a macroeconomic perspective, other literature, among which is Pissarides (2000, 2011), offers a ‘two-sided matching view’ of labour market equilibrium. However, this literature does not deal with the ‘who matches with whom’ problem; instead, using an aggregate matching function, determines the ‘total number’ of matches.
observed cell frequencies and the cell frequencies expected under the independence hypothesis). However, the usual approach in contingency tables is mostly global (chi-squared tests of independence, etc.), whereas in our paper we make an individual analysis for each cell, searching in particular which are the column (jobs) categories with greater propensity to match with each particular row (workers) category. Furthermore, in our paper we use the ratio between observed and expected cell frequencies instead of their difference.

The second major point addressed, deals with the dimensions problem generated when we define the row and column (worker and job) categories combining categories of several variables (for example sex, age, location, occupation, etc.). If more variables are combined, the number of combined categories may be very high and the large number of rows, columns and cells in the contingency table makes it difficult to get an overall picture. In addition, the number of cells in the table can be so high that, even with a large number of observations, in many cells the observed frequency may be zero, although, actually, its probability would not be zero (‘zero frequency problem’). Very low observed frequencies may also be problematic in order to estimate the corresponding probabilities. To solve this problem we use two approaches: clustering and factor decomposition of the multiple interactions among the different variables that are combined.

In a wide variety of fields, given a set of elements, clustering (or cluster analysis) is used to identify groups or ‘clusters’ of similar elements. Clustering is typically based on a measure of similarity. Among other things, it serves to reduce a large number of elements, which are our object of study, to a smaller and more manageable number of clusters. Biclustering (also known as block clustering, co-clustering, two-way clustering, etc.) is the simultaneous clustering of the rows and columns of a matrix such that the biclusters (or blocks) induced by the row/column partitions are homogeneous.

In this paper we apply the biclustering procedures to the contingency table which shows in each cell, as we said before, the observed number of matches between each row (workers) category and each column (jobs) category. We use the same measure of similarity proposed in Álvarez de Toledo et al. (2014), considering, in a matching context, that row (column) categories are more similar, the more they resemble in the way they match with column (row) categories. For example, in the application to the Spanish labour market in section 3, we

---

3 Among the general references in the voluminous clustering literature are Romesburg (2004), Xu and Wunsch (2008) and Everitt et al. (2011). References about biclustering are generally more specific. Govaert and Nadif (2014) offer an overview on the subject. Bock (2003) considers the biclustering process when the number of rows, columns and cells in the contingency table is very high.
consider that two categories of workers are more similar, the more they resemble in the way they match with the different categories of jobs and vice versa. We show in subsection 2.3 that this measure of similarity can be related to the Manhattan or City Block distance metric. Biclustering contributes to solve the aforementioned dimensions problem reducing the number of rows, columns and cells in the contingency table and increasing cell frequencies.

We use also another approach to this problem by decomposing the total effect of combined categories into partial effects of each category separately. Using a ‘Cobb-Douglas’ (or Log-Log) functional form we estimate the total propensity to match between each row combined category and each column combined category as a multiplicative function of the partial propensities to match between row individual categories and column individual categories. In this way we disentangle the mix of the effects of the multiple interactions between the different variables. We make an analogous analysis for similarities. Finally, we also clarify the cluster analysis by relating the clusters and biclusters obtained with combined variables to the clusters and biclusters obtained for each variable separately.

The rest of the paper is organised as follows. Section 2 is methodological. Subsection 2.1 presents the characteristics of the contingency tables that are the subject of study. Subsection 2.2 develops the concept of propensity to match and establishes the relation between the total propensity to match in each cell and the partial propensities to match or combine. Subsection 2.3 explains our clustering methodology and establishes the relation between the similarities and clusters for combined variables and for each variable separately. Section 3 applies the methodology to the Spanish labour market. Subsection 3.1 presents the data and subsections 3.2 and 3.3 are the application of the previous subsections 2.2 and 2.3 respectively. Finally, section 4 summarises the main conclusions of our research.

2. Methodology

2.1. Contingency tables

Let \( X \) and \( Y \) be two categorical variables with the same \( R \) categories each one, that are cross-classified in a \( R \times R \) square contingency table, and \( n_{ij} \) be the observed frequency in a random sample from a given population, for cell \((i, j), i, j = 1, ..., R\). The table is shown as Table 1. The row and column totals for the contingency table represent the marginal frequency distributions and \( n \) is the total sample size.
Let us suppose that the table also has the following characteristics:

a) The variables considered are mainly categorical. Quantitative variables (age, for example) are treated as if they were categorical, dividing its variation range in a limited number of intervals.

b) The order in which the categories are displayed is, in principle, arbitrary, although it is the same for rows and columns. Cluster analysis, as will be shown, can help to establish an order of proximity between categories.

c) Categories of $X$ (in rows) and $Y$ (in columns) are obtained through the combination of categories of different variables. There is a one-to-one correspondence, between rows and columns, of the variables combined and their categories. If we combine in rows $m$ variables ($v_1$ with $r_1$ categories, $v_2$ with $r_2$ categories, ... , $v_k$ with $r_k$ categories, ... , $v_m$ with $r_m$ categories), the total number of categories is

$$R = \prod_{k=1}^{m} r_k$$

(1)

As more variables are combined, the number of combined categories may be much higher. In columns there will be $m$ corresponding variables ($v_1$, $v_2$, ... , $v_k$, ... , $v_m$), with the same number of categories $r_k$ for each variable $v_k$ and the same total number of categories $R$.5

Each row $i = \{i_1, i_2, \ldots, i_k, \ldots, i_m\}$ corresponds to the combination of the particular category $i_1$ of $v_1$ with the particular category $i_2$ of $v_2$, etc. Each column $j = \{j_1, j_2, \ldots, j_k, \ldots, j_m\}$

---

5 For example, in our application to the Spanish labour market, we combine 52 categories of geographical location with 10 categories of occupation, resulting in 520 (52 x 10) combined categories. The categories in rows correspond to workers who match with the corresponding categories of jobs in columns.
corresponds to the combination of the particular category \( j_1 \) of \( v_1 \)' with the particular category \( j_2 \) of \( v_2 \)', etc. Each cell \((i, j)\) corresponds to the match of combined categories from both sides of the table: \( i = \{i_1, i_2, \ldots, i_k, \ldots, i_m\} \) in rows side, matches with \( j = \{ j_1, j_2, \ldots, j_k, \ldots, j_m\} \) in columns side, with an observed frequency \( n_{ij} \) and marginal totals \( n_{i+} \) and \( n_{+j} \).

The total sample size must be equal to

\[
 n = \sum_{i \in V_i} \sum_{j \in V_j} n_{ij} = \sum_{i \in V_i} n_{i+} = \sum_{j \in V_j} n_{+j} \tag{2}
\]

d) As a consequence of c), the number of cells in the table, \( R^2 \), can be so high that, even with a large number of observations, in many cells the observed frequency may be zero, although, actually, its probability is not zero (‘zero frequency problem’). Very low observed frequencies may also be problematic at the time of estimating the corresponding probabilities.

e) The large number of rows, columns and cells makes it difficult to get an overall picture (dimensions problem). Again, the decomposition of the matching process into partial components and the cluster analysis, which are discussed respectively in subsections 2.2 and 2.3, may be useful to solve this problem.

f) On the main diagonal each category in rows matches with its corresponding category in columns \((i_1 = j_1, i_2 = j_2, \ldots, i_k = j_k, \ldots, i_m = j_m)\). Frequencies on the main diagonal can be high, showing a high matching of each category with itself; although, a priori, it does not have necessarily to be so.

g) Data on observed matches in one-to-one two-sided matching markets (‘who matches with whom’), cross-classified by some characteristics of the parties in the match, can be considered a particular case of this type of contingency table (if the variables considered in the two sides of the market and their categories are the same). For example, in section 3 we will make an application to the Spanish labour market considering for the two sides, workers and jobs, the variables geographical location and occupation.

---

6 That is, a particular category of the variable \( X \) (in rows) matches with the same particular category of the variable \( Y \) (in columns). For example, in our application to the Spanish labour market, these are matches of workers of a given province and occupation group with jobs corresponding to the same province and occupation group (therefore, without geographical and occupational mobility).
In addition to the observed frequencies in each cell and marginal totals, we will use the totals corresponding to each of the categories of each variable separately. That is, the total frequency of each category $i_k$ of $v_k$ in rows

$$n_{i_k} = \sum_{i_k \in i} n_{i+}$$  \hspace{1cm} (3)

($n_{i+}$ will be obtained by adding the marginal totals $n_{i+}$ for all the combined categories $i$ in which appears $i_k$) and the total frequency of each category $j_k$ of $v_k$ in columns

$$n_{+j_k} = \sum_{j_k \in j} n_{+j}$$  \hspace{1cm} (4)

and therefore

$$n = \sum_{v_{ik}} n_{i_k} = \sum_{v_{jk}} n_{j_k}$$  \hspace{1cm} (5)

Finally, we will use also the frequency with which each category $i_k$ of each variable $v_k$ on the side of the rows, matches with each category $j_k$ of the corresponding variable $v_k$’ on the side of the columns

$$n_{i_kj_k} = \sum_{i_k \in i \land j_k \in j} n_{ij}$$  \hspace{1cm} (6)

and therefore

$$n = \sum_{v_{ikj_k}} n_{i_kj_k}$$  \hspace{1cm} (7)

The relative frequencies give us estimates of four kinds of probabilities:

- the joint probability mass function (empirical probabilities)

$$p_{ij} = \frac{n_{ij}}{n}$$  \hspace{1cm} (8)

- the marginal probability mass functions

$$p_{i+} = \frac{n_{i+}}{n}$$  \hspace{1cm} (9)

$$p_{+j} = \frac{n_{+j}}{n}$$  \hspace{1cm} (10)
the separate probability of each of the categories of each variable $v_k$ in rows and each variable $v_k'$ in columns, or \textit{individual probabilities}

\begin{align*}
p_{ik+} &= \frac{n_{ik+}}{n} \\
p_{+jk} &= \frac{n_{+jk}}{n}
\end{align*}  \hspace{1cm} (11)

- and the probability that each category of each variable on the side of the rows matches with each category of the corresponding variable on the side of the columns, or \textit{individual matching probabilities}

\begin{align*}
p_{ikj_k} &= \frac{n_{ikj_k}}{n} 
\end{align*}  \hspace{1cm} (13)

In subsection 2.3 we will use also the \textit{conditional probabilities} of an event of category $j$ (in columns) given an event of category $i$ (in rows)

\begin{align*}
p(j|i) &= \frac{p(j \cap i)}{p(i)} = \frac{p_{ij}}{p_{i+}}
\end{align*}  \hspace{1cm} (14)

and of an event of category $i$ (in rows) given an event of category $j$ (in columns)

\begin{align*}
p(i|j) &= \frac{p(i \cap j)}{p(j)} = \frac{p_{ij}}{p_{+j}}
\end{align*}  \hspace{1cm} (15)

The conditional probabilities in (14) and (15) are also commonly referred to as \textit{row and column proportions} respectively, and also as \textit{row and column profiles} (the latter are often used in connection with graphical displays).

2.2. Matching

Consider the hypothesis of independence between $X$ and $Y$. In this case the joint probability would be the product of the corresponding marginal probabilities

\begin{align*}
p_{ij} &= p_{i+} p_{+j} 
\end{align*}  \hspace{1cm} (16)

and also

\begin{align*}
p_{ikj_k} &= p_{ik+} p_{+jk} 
\end{align*}  \hspace{1cm} (17)
Additionally, consider the hypothesis of independence between the $m$ variables combined in rows and in columns. Then, the probability of each combination would be the product of the probabilities of their components

$$p_{i+} = \prod_{k=1}^{m} p_{ik+}$$  

$$p_{+j} = \prod_{k=1}^{m} p_{+jk}$$

In what follows, we will denote these probabilities obtained under both hypotheses of independence, \textit{random} probabilities (denoted as $pr$), because we would get them if we randomly combine and match each category of each variable, given their probabilities separately.

$$pr_{i+} = \prod_{k=1}^{m} p_{ik+}$$  

$$pr_{+j} = \prod_{k=1}^{m} p_{+jk}$$

$$pr_{ij} = pr_{i+} pr_{+j}$$

$$pr_{ikj} = p_{ik+} p_{+jk}$$

We can measure the \textit{total propensity to match} between the row and column combinations in cell $(i, j)$ as the ratio of the probability estimated from the observed frequencies to the random probability

$$pm_{ij} = \frac{p_{ij}}{pr_{ij}}$$

Values higher than one mean that the propensity is greater than in the random case, and vice versa. Moreover, it is straightforward that, by using $pr_{ij}$ as weights, the weighted mean of the $pm_{ij}$ values in all the cells is one. If we take as given the row and column combinations of categories and focus exclusively in the row-column matching, we can define alternatively the \textit{row-column propensity to match} as

$$pm'_{ij} = \frac{p_{ij}}{p_{i+} p_{+j}}$$
Due to the dimensions problem outlined in subsection 2.1, the analysis of the total propensity to match in each cell may require the consideration of a very large number of cases. Furthermore, in each case there is a mix of effects from the multiple interactions between the different variables on both sides. We will quantify separately these effects using a much smaller number of partial propensities to match or combine:

- propensity to match between each category of each variable on the side of the rows, and each category of the corresponding variable on the side of the columns\(^7\), or partial propensities to match

\[
dm_{ikjk} = \frac{p_{ikjk}}{pr_{ikjk}}
\]

- propensity of the categories of the different variables on the side of the rows to combine between them

\[
dm_{i+} = \frac{p_{i+}}{pr_{i+}}
\]

- propensity of the categories of the different variables on the side of the columns to combine between them

\[
dm_{+j} = \frac{p_{+j}}{pr_{+j}}
\]

From the above equations, we can relate the observed frequencies to the propensities to match or combine, the individual probabilities and the total sample size

\[
n_{ij} = n p_{ij} = n pr_{ij} dm_{ij} = n \left( \prod_{k=1}^{m} p_{ik+} \right) \left( \prod_{k=1}^{m} p_{+jk} \right) dm_{ij}
\]

\[
n_{i+} = n p_{i+} = n pr_{i+} dm_{i+} = n \left( \prod_{k=1}^{m} p_{ik+} \right) dm_{i+}
\]

\[
n_{+j} = n p_{+j} = n pr_{+j} dm_{+j} = n \left( \prod_{k=1}^{m} p_{+jk} \right) dm_{+j}
\]

Also, from (22), (24), (25), (27) and (28), we obtain the relation between the total propensity to match, the row-column propensity to match and the propensities of the categories of the different variables, on the side of the rows and on the side of the columns, to combine between them

\(^7\) We have not considered the propensity to match of each category of each variable on the side of the rows, with each category of other variables on the side of the columns because, at least in the cases we have analysed, it does not seem relevant.
\[ pm_{ij} = pm'_{ij} pm_{i+} pm_{+j} \]  

(32)

As can be seen, the total propensity to match includes the effects of the other three propensities.

The set of total propensities in each cell which corresponds to a given set of partial propensities is not univocally determined, but, on the contrary, there are infinite solutions. Bearing this in mind, for a given sample data, we can try to specify a regression equation that explains as much as possible the total propensities in terms of the partial propensities. From equation (29), we regress the equation

\[ n_{ij} = n \left( \prod_{k=1}^{m} p_{ik+} \right) \left( \prod_{k=1}^{m} p_{+jk} \right) \hat{pm}_{ij} + \varepsilon \]  

(33)

where \( \hat{pm}_{ij} \) is the estimated value of \( pm_{ij} \)

\[ \hat{pm}_{ij} = \left( \prod_{k=1}^{m} pm_{ik+jk}^{\alpha_k} \right) pm_{i+} pm_{+j} \]  

(34)

\( \hat{n}_{ij} \) is the estimated value of \( n_{ij} \)

\[ \hat{n}_{ij} = n \left( \prod_{k=1}^{m} p_{ik+} \right) \left( \prod_{k=1}^{m} p_{+jk} \right) \hat{pm}_{ij} \]  

(35)

\( \alpha_k, \beta \) and \( \gamma \) are parameters to be estimated and \( \varepsilon \) is an error term.

The use of the ‘Cobb-Douglas’ functional form in (34) reflects the idea that the effects of the partial propensities multiply each other (rather than add to each other). In the application to the Spanish labour market that we will show in section 3, we obtain a good fit in the regression, \( \alpha_k \approx 1 \) and \( \beta + \gamma \approx 1 \).

The equations in factorial form (34) and (35) can be completed with a residual factor \( rf \), calculated as the ratio of the observed value to the estimated value\(^8\). From (29) and (35)

\[ rf = \frac{n_{ij}}{\hat{n}_{ij}} = \frac{pm_{ij}}{\hat{pm}_{ij}} \]  

(36)

\( \hat{n}_{ij} \)

---

\(^8\) From (33), (36) and (38) we obtain the relation between the additive error term \( \varepsilon \) and the multiplicative residual factor \( rf \)

\[ \varepsilon = n_{ij} - \hat{n}_{ij} \]

\[ rf = n_{ij} / \hat{n}_{ij} \]

and therefore

\[ \varepsilon = (rf - 1) \hat{n}_{ij} \]
and

\[ pm_{ij} = \hat{pm}_{ij} \, rf = \left( \prod_{k=1}^{m} p_{ik+jk}^{\alpha_k} \right) \left( \prod_{k=1}^{m} p_{i+kj}^{\beta} \right) \left( \prod_{k=1}^{m} p_{+j+k}^{\gamma} \right) \, rf \]  

From (37) and (32) we also obtain the factor decomposition of the row-column propensity to match

\[ pm'_{ij} = \left( \prod_{k=1}^{m} p_{ik+jk}^{\alpha_k} \right) \left( \prod_{k=1}^{m} p_{i+kj}^{\beta} \right) \left( \prod_{k=1}^{m} p_{+j+k}^{\gamma} \right) \, rf \]  

The factorial equation (37) allows us to quantify separately the contributions of each of the partial propensities to the total propensity to match in each cell, as well as the contribution of the residual factor (remember that values higher than one imply that propensities are greater than in the random case, and vice versa). The residual factor measures the part of the total propensity not explained by the partial propensities. For example, in the application to the Spanish labour market in section 3, we will observe that the contribution of partial propensities to match between geographical locations of workers and jobs is greater than that of partial propensities to match between occupations, and both \((pm_{i+jk})\) are greater than that of the propensities to combine \(pm_{i+k}\) and \(pm_{+j}\) (between geographical locations and occupations of workers, or between geographical locations and occupations of jobs). Also, we will show that the residual factor increases with lower values of observed frequencies, which makes the estimation of equation (37) less accurate, but it may reflect also more complex interactions than the corresponding to the partial propensities. Furthermore, each of the factors can be analysed separately, examining which cases correspond to the highest and lowest values of the partial propensities (for instance, which are the locations of workers and jobs with greater propensity to match).

Finally, when, as we have seen in subsection 2.1, empirical probabilities are not reliable because they are based on zero or very low observed frequencies, we can use alternatively the frequencies estimated from equation (35).

2.3. Clustering

As we have presented in the previous sections, we consider a square \(R \times R\) contingency table in which the total number of categories \(R\) can be very large and, as a consequence, the number of cells in the table, \(R^2\), can be so high that, even with a large number of observations, in
many cells the observed frequency may be zero or very low, and therefore the empirical probabilities based on them are not reliable. Moreover, the order in which the categories are displayed is, in principle, arbitrary. Finally, the large number of rows, columns and cells makes it difficult to get an overall picture of the object to which the contingency table refers. In particular, a comprehensive analysis of the total propensities to match (‘who matches with whom’) in each cell may be difficult if we must consider a very large number of cases. Clustering methodology permits to address all these problems ordering and grouping the \( R \) categories (in rows and columns) in a lower number of clusters and the \( R \times R \) cells in a lower number of biclusters.

The clustering methodology is usually based on a previously defined similarity (or dissimilarity) measure between the elements that are clustered. In a matching context, we consider that row (column) categories are more similar, the more they resemble in the way they match with column (row) categories. For instance, in the application to the Spanish labour market in section 3, we consider that two categories of workers are more similar, the more they resemble in the way they match with the different categories of jobs and vice versa. Then, we measure the similarity between each pair of categories \( i_A-i_B \) (rows of the contingency table) as the overlapping or percentage of coincidence of their row profiles (distribution of their conditional probabilities \( p_{ij}/p_{i+} \) of matching with each of the different column categories \( j \)).

\[
sim_{i_A-i_B} = \sum_j \min \left( \frac{p_{i_Aj}}{p_{i_A+}} , \frac{p_{i_Bj}}{p_{i_B+}} \right)
\]

or also, taking into account (25)

\[
sim_{i_A-i_B} = \sum_j p_{+j} \min \left( p_{m{i_Aj}}' / p_{m{i_Bj}}' \right)
\]

This measure of similarity can be related to the Manhattan or City Block distance metric

\[
dist_{i_A-i_B} = \sum_j \left| \frac{p_{i_Aj}}{p_{i_A+}} - \frac{p_{i_Bj}}{p_{i_B+}} \right|
\]

with

\[
sim_{i_A-i_B} = 1 - 0.5 \, \text{dist}_{i_A-i_B}
\]
Its value will be between one (if the row profiles are identical) and zero (if their intersection is null). We can measure the similarity between each pair of categories $j_A j_B$ (columns of the contingency table) in an analogous way.

\[
sim_{j_A j_B} = \sum_i \min \left( \frac{p_{ij_A}}{p_{+j_A}}, \frac{p_{ij_B}}{p_{+j_B}} \right) = \sum_i p_{i+} \min \left( p_{m'_{ij_A}}, p_{m'_{ij_B}} \right)
\]

(44)

Remember that the category of $X$ or row of the contingency table $i_A = \{i_{1A}, i_{2A}, \ldots, i_{kA}, \ldots, i_{mA}\}$ corresponds to the combination of the particular categories $i_{kA}$ of each of the $m$ variables $v_k (k = 1, 2, \ldots, m)$ and that $r_k$ is the number of different categories of $v_k (i_k = 1, 2, \ldots, r_k)$. The same applies to $i_B$, $j_A$ and $j_B$. From (41) and (44) we can deduce that similarity will be high between rows $i_A i_B$ (columns $j_A j_B$) with similar values of $pm'_{iA}$ and $pm'_{iB}$ ($pm'_{ij_A}$ and $pm'_{ij_B}$), which, in turn, implies similar factors in (39). As we will observe in our application, partial propensities to match $pm_{iA jB}$ usually play a major role in this equation. Therefore $sim_{iA-iB}$ will be high if $pm_{iA jB} \approx pm_{iB jB}$ for each of the $m$ variables ($k = 1, 2, \ldots, m$) and each of the $r_k$ categories of $j_k (j_k = 1, 2, \ldots, r_k)$. For example, in our application to the Spanish labour market, we consider for the two sides, workers and jobs, the variables geographical location and occupation. In this case, similarity between the two categories of workers (location-occupation)$_A$ and (location-occupation)$_B$ will be high if the partial propensities to match with each location of jobs are similar for worker’s location$_A$ and worker’s location$_B$ and the partial propensities to match with each occupation of jobs are similar for worker’s occupation$_A$ and worker’s occupation$_B$. Moreover, $sim_{iA-iB}$ will be even higher if the categories of some variables are coincident (for instance, the same location combined with two different occupations or vice versa). A similar argument applies to $sim_{jA-jB}$.

Again, due to the dimensions problem, the detailed analysis of similarities between each pair of rows or columns may require the consideration of a very large number of cases and, in each case, there is a mix of effects from the multiple interactions between the different variables on both sides. We will quantify separately these effects using a much smaller number of partial similarities for each of the $m$ variables ($k = 1, 2, \ldots, m$):

---

9 Not surprisingly, we will find that this occurs if worker’s location$_A$ and worker’s location$_B$ are close geographically.
- partial similarities between each pair of categories $i_{kA} - i_{kB}$ of each variable $v_k$ (in the rows of the contingency table) measured as the percentage of coincidence of their row profiles (considering only the variable $v_k$)

$$sim_{i_{kA} - i_{kB}} = \sum_{j_k} \min \left( \frac{p_{i_{kA}j_k}}{p_{i_{kA}+}}, \frac{p_{i_{kB}j_k}}{p_{i_{kB}+}} \right) = \sum_{j_k} p_{+j_k} \min \left( pm_{i_{kA}j_k}^{+}, pm_{i_{kB}j_k}^{+} \right)$$  \hspace{1cm} (45)

- partial similarities between each pair of categories $j_{kA} - j_{kB}$ of each variable $v_k$ (in the columns of the contingency table) measured as the percentage of coincidence of their column profiles (considering only the variable $v_k$)

$$sim_{j_{kA} - j_{kB}} = \sum_{i_k} \min \left( \frac{p_{i_{kA}j_k}}{p_{+i_{kA}}}, \frac{p_{i_{kB}j_k}}{p_{+i_{kB}}} \right) = \sum_{i_k} p_{i_k+} \min \left( pm_{i_{kA}j_k}^{+}, pm_{i_{kB}j_k}^{+} \right)$$  \hspace{1cm} (46)

The values of partial similarities are also between one (if the row profiles are identical) and zero (if their intersection is null). They will be high if the partial propensities to match in (45) or in (46) are similar for each of the $r_i$ categories of $j_k$ and $i_k$, respectively.

As with the propensities, the set of similarities $sim_{i_{kA} - i_B}$ (or $sim_{j_{kA} - j_B}$) which corresponds to a given set of partial similarities $sim_{i_{kA} - ikB}$ (or $sim_{j_{kA} - jkB}$), is not univocally determined, but, on the contrary, there are infinite solutions. Equations (39), (41), (44), (45) and (46) would allow relating them but, however, this relation is too complicated. Instead, just as we make with propensities, for a given sample data, we can try to specify a regression equation that explains as much as possible the similarities $sim_{i_{kA} - iB}$ in terms of the partial similarities $sim_{i_{kA} - ikB}$ (or the similarities $sim_{j_{kA} - jB}$ in terms of the partial similarities $sim_{j_{kA} - jkB}$). In our application to the Spanish labour market in section 3, we have obtained the best results with the ‘Cobb-Douglas’ functional form

$$sim_{i_{kA} - iB} = \prod_{k=1}^{m} sim_{i_{kA} - ikB}^{\alpha_k} + \varepsilon$$  \hspace{1cm} (47)

$$sim_{j_{kA} - jB} = \prod_{k=1}^{m} sim_{j_{kA} - jkB}^{\alpha_k} + \varepsilon$$  \hspace{1cm} (48)

$\alpha_k$ are parameters to be estimated and $\varepsilon$ is an error term (the two equations have the same form but they are estimated with different data, so the estimates of $\alpha_k$ will be different).
As Álvarez de Toledo et al. (2014) show, in a square contingency table, with corresponding categories in rows and columns, the similarities \( \text{sim}_{i_A-i_B} \) between each pair of categories \( i_A-i_B \) (in rows) are highly positively correlated with the similarities \( \text{sim}_{j_A-j_B} \) between the corresponding pair of categories \( j_A-j_B \) (in columns). The same applies for \( \text{sim}_{i_kA-i_kB} \) and \( \text{sim}_{j_kA-j_kB} \). As we will see in the application, it may be convenient to use the same measure of similarity in the clustering process of rows and columns in order to obtain the same grouping on both sides. In this case, we will use the arithmetic mean of the above expressions for rows and columns.

\[
\text{sim}_{c_A-c_B} = \frac{\left( \text{sim}_{i_A-i_B} + \text{sim}_{j_A-j_B} \right)}{2} \tag{49}
\]

\[
\text{sim}_{c_kA-c_kB} = \frac{\left( \text{sim}_{i_kA-i_kB} + \text{sim}_{j_kA-j_kB} \right)}{2} \tag{50}
\]

where \( c_A-c_B \) means the corresponding pair of combined categories in rows \( (i_A-i_B) \) and in columns \( (j_A-j_B) \), and \( c_kA-c_kB \) the corresponding pair of categories of each variable \( v_k \) in rows \( (i_kA-i_kB) \) and in columns \( (j_kA-j_kB) \).

And, instead of (47) and (48), we will estimate

\[
\text{sim}_{c_A-c_B} = \prod_{k=1}^{m} \text{sim}_{c_kA-c_kB}^{\alpha_k} + \varepsilon \tag{51}
\]

In the application to the Spanish labour market (section 3), we obtain an acceptable fit in the regression and \( \alpha_k \approx 1 \).

Again, equations (47), (48) and (51) can be completed with a residual factor \( rf \), calculated as the ratio of the observed value to the estimated value. The equations allow us to analyse to what extent high and low values of partial similarities contribute to high and low values of the similarity of each pair of combined categories, as well as the contribution of the residual factor.

Based on such former similarity measures, we use a hierarchical method of clustering, merging the two categories with the highest similarity into a new category and going on successively, with categories gradually fusing to form increasingly larger categories or clusters. At the same time we can get the corresponding contingency tables with a decreasing number of clusters (rows and columns) and biclusters (combinations of a row cluster and a
column cluster). Also, the order in which the categories are displayed (order that was, in principle, arbitrary) is changed so that the categories with greater similarity are closer.

Using the same measure of similarity in the clustering process of rows and columns, we get the same grouping and ordering on both sides and, as shown in Álvarez de Toledo et al. (2014), we obtain a main diagonal of biclusters with (usually) high propensity to match. For example, in our application to the Spanish labour market, groups of categories location-occupation of workers and jobs with high similarity and propensity to match between them. It is also interesting the analysis of high propensities to match out of the main diagonal possibly corresponding to isolated specific cases (for instance a particular combination location-occupation of workers with high propensity to match with a particular combination location-occupation of jobs). Instead of the propensity to match for each of the cells of the original contingency table, now we can calculate the row-column propensity to match for each bicluster\textsuperscript{10}.

\textbf{Figure 1. Contingency table, represented as a ‘matching map’.}
\textit{(A darker cell represents higher propensity)}

In Fig. 1 we represent the contingency table as a ‘matching map’, showing in each cell or bicluster the corresponding propensity to match. We can use different levels of ‘zoom’ depending on whether you want to have an overview of the whole map or a more detailed view of a particular area\textsuperscript{11}.

\textsuperscript{10} The row-column propensity to match for each bicluster will be equal to the weighted average of the row-column propensities to match of all cells in the bicluster. The weight of each cell in the bicluster is the product of the weight of its row marginal total frequency and the weight of its column marginal total frequency.

\textsuperscript{11} An example of a matching map, applied to the Andalusian labour market, can be found in Álvarez de Toledo et al. (2013), although with different data from ours and with a different (simpler) measure of similarity, which is based on correlations.
Once more, the detailed analysis of clusters and biclusters may require the consideration of a very large number of cases and, in each case, there is a mix of effects from the multiple interactions between the different variables. The analysis can be clarified by relating it to the clusters obtained for each variable separately. Given the correspondence between clusters and similarities, and the relation (51) between the similarities of combined categories of different variables and the partial similarities of each variable, an analogous relation can be obtained between the clusters for combined variables and for each variable separately. For instance, in our application to the Spanish labour market, we will find that clusters for location and occupation combined are closely related to clusters for location and occupation separately.

3. Application to the Spanish labour market

3.1. Data

In our application, the two categorical variables that are cross-classified in a square contingency table are characteristics of workers (rows) and jobs (columns) in a random sample of matches between them. The contingency table is shown as Table 2. R = 520 combined categories (in rows and columns) are formed by combining \( r_1 = 52 \) categories of the variable location with \( r_2 = 10 \) categories of the variable occupation. Each row \( i = \{ i_1, i_2 \} \) corresponds to the combination of the particular category \( i_1 \) of the variable worker location \( v_1 \) with the particular category \( i_2 \) of the variable worker occupation \( v_2 \), and each column \( j = \{ j_1, j_2 \} \) corresponds to the combination of the particular category \( j_1 \) of the variable job location \( v_1' \) with the particular category \( j_2 \) of the variable job occupation \( v_2' \). The data used come from the MCVL (Muestra Continua de Vidas Laborales or Continuous Working Life Sample), an annual sample taken from administrative data of the Social Security system in Spain consisting of more than one million individuals across the country (4% of all individuals affiliated to the Social Security system). Our data correspond to \( n = 1,967,441 \) job matches (new contracts) during the period 2011-2013. The 52 categories of the variable location are the 50 Spanish provinces and the 2 Autonomous Cities (Ceuta and Melilla) to which we will refer simply as ‘52 provinces’ in what follows. For workers, location corresponds to their place of habitual residence, while for jobs, it corresponds to the place where the main activity of the firm is located. For jobs, the 10 categories of the variable occupation correspond to the main contribution groups to the Social Security, while for workers they correspond to the contribution groups of their former job. The order in which the provinces and contribution groups are displayed is, in principle, arbitrary, although it is the same for rows and columns. Later, clustering will establish an order of proximity between categories.
Table 2. Contingency table for worker-job matches.

<table>
<thead>
<tr>
<th>Worker categories</th>
<th>Location 1</th>
<th>Location 2</th>
<th>...</th>
<th>Location 10</th>
<th>Location 11</th>
<th>...</th>
<th>Location 20</th>
<th>...</th>
<th>Location j</th>
<th>...</th>
<th>Location 511</th>
<th>...</th>
<th>Location 520</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job categories</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>10</td>
<td>11</td>
<td>...</td>
<td>20</td>
<td>...</td>
<td>j</td>
<td>...</td>
<td>511</td>
<td>...</td>
<td>520</td>
<td>n</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>10</td>
<td>11</td>
<td>...</td>
<td>20</td>
<td>...</td>
<td>j</td>
<td>...</td>
<td>511</td>
<td>...</td>
<td>520</td>
<td>n</td>
</tr>
</tbody>
</table>

In what follows, it will be convenient to specify clearly the combinations location-occupation for workers and jobs matches. Thus, \( n_{ij} = n_{i1j1}, n_{ij2} \) is the frequency of matches of workers of location \( i_1 \) and occupation \( i_2 \) with jobs of location \( j_1 \) and occupation \( j_2 \), \( n_{i1} \) is the frequency of matches of workers of location \( i_1 \), \( n_{i2} \) is the frequency of matches of workers of occupation \( i_2 \), \( n_{j1} \) is the frequency of matches of jobs of location \( j_1 \), \( n_{j2} \) is the frequency of matches of jobs of occupations \( j_2 \). Finally, \( n_{i1j1} \) is the frequency of matches of workers of location \( i_1 \) with jobs of location \( j_1 \), \( n_{i2j2} \) is the frequency of matches of workers of occupation \( i_2 \) with jobs of occupation \( j_2 \), \( n_{i1} \) is the frequency of matches of workers which combine location \( i_1 \) with occupation \( i_2 \) and \( n_{j1j2} \) is the frequency of matches of jobs which combine location \( j_1 \) with occupation \( j_2 \).

Although we have a large number of observations (\( n = 1,967,441 \)), as the number of cells in the table (\( R \times R = 520 \times 520 = 270,400 \)) is high and, also, the distribution of frequencies is very unequal, in many cells (86%) the observed frequency is zero, although actually its probability may not be zero (‘zero frequency problem’). Also, we have a large number of very low frequencies (for example, 43% of the non-zero frequencies are 1) which can be problematic when estimating the corresponding probabilities.

The large number of rows, columns and cells makes it difficult to get an overall picture of the matching process (dimensions problem). In the next sections we will use cluster analysis, as well as the factor decomposition of propensities to match, to solve this problem.

Frequencies on the main diagonal are very high, which show a very high matching of each category location-occupation (of workers) with the same category location-occupation (of
jobs). This means that the mobility (geographical and occupational) in the labour market is low. In 67% of the matches there is no change in the location (province) nor the occupation (contribution group), in 20% of the matches there are changes in the occupation but not in the location, in 10% of the matches there are changes in the location but not in the occupation, and only in 3% of the matches there are changes in the location and the occupation.

3.2. Matching

In our application, the total propensity to match between workers of location \(i_1\) and occupation \(i_2\) with jobs of location \(j_1\) and occupation \(j_2\) is

\[
p_{m_{ij}} = p_{m_{i_1j_2j_1j_2}} = \frac{p_{i_1i_2j_1j_2}}{p_{r_{i_1i_2j_1j_2}}} \tag{52}
\]

where \(p_{i_1i_2j_1j_2}\) is the empirical probability of matches of the type \(i_1i_2j_1j_2\), derived from the observed frequencies

\[
p_{i_1i_2j_1j_2} = \frac{n_{i_1i_2j_1j_2}}{n} \tag{53}
\]

and

\[
p_{r_{i_1i_2j_1j_2}} = p_{i_1}p_{i_2}p_{j_1}p_{j_2} \tag{54}
\]

with individual probabilities

\[
p_{i_1} = \frac{n_{i_1}}{n}; \quad p_{i_2} = \frac{n_{i_2}}{n}; \quad p_{j_1} = \frac{n_{j_1}}{n}; \quad p_{j_2} = \frac{n_{j_2}}{n} \tag{55}
\]

is the random probability of matches of the type \(i_1i_2j_1j_2\), under the hypothesis of independence between all the variables. We denote these probabilities as ‘random’ because, given the individual probabilities, we would get them if we randomly match workers and jobs, and also combine randomly locations and occupations of workers and jobs.

In subsection 3.3 we will use also the row-column propensity to match between workers of location \(i_1\) and occupation \(i_2\) with jobs of location \(j_1\) and occupation \(j_2\)

\[
p_{m'_{ij}} = p_{m'_{i_1i_2j_1j_2}} = \frac{p_{i_1i_2j_1j_2}}{p_{i_1}p_{j_1}p_{j_2}} = \frac{p_{m_{i_1i_2j_1j_2}}}{p_{m_{i_1i_2}p_{j_1j_2}}} \tag{56}
\]

where

\[
p_{i_1i_2} = \frac{n_{i_1i_2}}{n}; \quad p_{j_1j_2} = \frac{n_{j_1j_2}}{n} \tag{57}
\]
Due to the high number of cells in the table, the analysis of the total propensity to match in each cell may require the consideration of too many cases. Furthermore, in each case there is a mix of effects of the propensities to match between locations, the propensities to match between occupations and the propensities of locations and occupations to combine between them on the side of workers and on the side of jobs. We will quantify separately these effects using a much smaller number of partial propensities to match or combine, using the equation (37) that, in our application, takes the form

$$pm_{i_1i_2j_1j_2} = pm_{i_1j_1}^\alpha pm_{i_2j_2}^\beta pm_{i_1i_2}^\gamma pm_{j_1j_2}^\gamma rf$$

where $pm_{i_1j_1}$ is the propensity to match between workers of location $i_1$ and jobs of location $j_1$

$$pm_{i_1j_1} = \frac{p_{i_1j_1}}{(p_{i_1} p_{j_1})} \quad \text{where} \quad p_{i_1j_1} = \frac{n_{i_1j_1}}{n}$$

$pm_{i_2j_2}$ is the propensity to match between workers of occupation $i_2$ and jobs of occupation $j_2$

$$pm_{i_2j_2} = \frac{p_{i_2j_2}}{(p_{i_2} p_{j_2})} \quad \text{where} \quad p_{i_2j_2} = \frac{n_{i_2j_2}}{n}$$

$pm_{i_1i_2}$ is the propensity of workers to combine location $i_1$ with occupation $i_2$

$$pm_{i_1i_2} = \frac{p_{i_1i_2}}{(p_{i_1} p_{i_2})} \quad \text{where} \quad p_{i_1i_2} = \frac{n_{i_1i_2}}{n}$$

$pm_{j_1j_2}$ is the propensity of jobs to combine location $j_1$ with occupation $j_2$

$$pm_{j_1j_2} = \frac{p_{j_1j_2}}{(p_{j_1} p_{j_2})} \quad \text{where} \quad p_{j_1j_2} = \frac{n_{j_1j_2}}{n}$$

and $rf$ is the residual factor (the part of the total propensity not explained by the partial propensities).

To estimate the parameters $\alpha$, $\beta$, $\gamma$, we use equations (33) and (34), which, in our application, are equivalent to regressing

$$n_{i_1i_2j_1j_2} = \hat{n}_{i_1i_2j_1j_2} + \epsilon = n pr_{i_1i_2j_1j_2} \hat{pm}_{i_1i_2j_1j_2} + \epsilon$$

where $\hat{pm}_{i_1i_2j_1j_2}$ is the estimated value of $pm_{i_1i_2j_1j_2}$

$$\hat{pm}_{i_1i_2j_1j_2} = pm_{i_1j_1}^\alpha pm_{i_2j_2}^\beta pm_{i_1i_2}^\gamma pm_{j_1j_2}^\gamma$$

and $\hat{n}_{i_1i_2j_1j_2}$ is the estimated value of $n_{i_1i_2j_1j_2}$
\[ \hat{h}_{i_1i_2j_1j_2} = n \hat{p} t_{i_1i_2j_1j_2} \hat{p} m_{i_1i_2j_1j_2} \]  

(65)

As the variables \(pm_{i_1i_2}\) and \(pm_{j_1j_2}\) are highly positively correlated (\(\text{corr}(pm_{i_1i_2}, pm_{j_1j_2}) = 0.928\)), we estimate only one parameter \(\beta\) for the product of both variables. Equation (58), that we will use to analyse the total propensity to match as a function of the partial propensities, takes then the form

\[ pm_{i_1i_2j_1j_2} = pm_{i_1j_1}^{\alpha_1} pm_{i_2j_2}^{\alpha_2} (pm_{i_1i_2j_1j_2})^{\beta} rf \]

or, taking into account (56)

\[ pm_{i_1i_2j_1j_2} = pm_{i_1j_1}^{\alpha_1} pm_{i_2j_2}^{\alpha_2} (pm_{i_1i_2j_1j_2})^{\beta-1} rf \]

(66)

(67)

We show the results of the estimation of equation (66) in Table 3. We obtain a good fit in the regression, values of \(\alpha_1\) and \(\alpha_2\) very close to 1, and of \(\beta\) close to 0.5 (so that \((pm_{i_1i_2} pm_{j_1j_2})^{\beta}\) would be approximately the geometric mean of both variables).

| Table 3. Total propensity to match as a function of partial propensities. |
|-----------------|------|------|------|
|                 | \(\alpha_1\) | \(\alpha_2\) | \(\beta\) |
| Coefficients    | 1.01*** | 0.99*** | 0.55*** |
| Standard errors | (0.0003) | (0.0005) | (0.0006) |
| Number of observations | 38,468 |
| Adjusted R-squared | 0.99 |

* p<.1; ** p<.05; *** p<.01

On the basis of equation (66), we analyse first the partial propensities and, later, their contribution to the total propensity to match. Before that, we want to make two considerations:

Firstly, we observe that propensities are highly dependent of the type of mobility in the match:

a) no mobility \((i_1 = j_1; i_2 = j_2)\)

b) only occupational mobility \((i_1 = j_1; i_2 \neq j_2)\)

c) only geographical mobility \((i_1 \neq j_1; i_2 = j_2)\)

d) geographical and occupational mobility \((i_1 \neq j_1; i_2 \neq j_2)\).

Propensities decrease sharply from a) to d), reflecting low mobility.
Secondly, the total and partial propensities to match $pm_{i_1i_2j_1j_2}$, $pm_{i_1j_1}$ and $pm_{i_2j_2}$ show a high dependence of the ‘size’ of the corresponding cells measured as $p_{i_1}p_{i_2}p_{j_1}p_{j_2}$, $p_{i_1}p_{j_1}$ and $p_{i_2}p_{j_2}$, respectively (random probabilities of the cells, or corresponding percentage of the total sample size in the random case). We will show and explain in each case this dependence.

We analyse in the first place the propensity to match between workers and jobs from the same province ($pm_{i_1j_1}$, with no geographical mobility, $i_1 = j_1$). In the left side of Fig. 2 we represent the values of $pm_{i_1j_1}$ and the size of the corresponding cell for the 52 provinces.

![Figure 2. Partial propensity to match within and between provinces.](image)

The propensity to match within the same province is high (well above 1 with median = 59) for all the provinces, reflecting the low geographical mobility. Also, it can be seen that there is a very clear inverse relation between propensity to match and size. The simplest explanation of this is that the high proportion of matches without mobility does not vary much across provinces (83% on average). Then, if we consider that approximately

$$\frac{n_{i_1j_1}}{n_{i_1}} = \frac{p_{i_1j_1}}{p_{i_1}} = \frac{n_{i_1j_1}}{n_{j_1}} = \frac{p_{i_1j_1}}{p_{j_1}} = b$$

(68)

where $b$ is a constant, we get\(^{12}\)

$$pm_{i_1j_1} = \frac{p_{i_1j_1}}{(p_{i_1} p_{j_1})} = b(p_{i_1} p_{j_1})^{-1/2} = b(size)^{-1/2}$$

(69)

which fits quite well Fig. 2.

\(^{12}\) From equation (68)

$$p_{i_1j_1} = b p_i = b p_j = b(p_i p_j)^{1/2}$$
Now, we analyse the propensity to match between workers and jobs from different provinces \( (pm_{i_1j_1}, \text{with geographical mobility}, \ i_1 \neq j_1) \). On the right side of Fig. 2 we represent the values of \( pm_{i_1j_1} \) vs. kilometric distance between province capitals\(^{13}\).

The propensity to match between different provinces is much lower than in the previous case, reflecting the low geographical mobility. Only 5% of the values are greater than 1; median = 0.05. The figure also shows how the propensity decreases as the kilometric distance increases. Again, the size of the cell affects the propensity, but the influence is relatively small\(^{14}\). In Table 4 we show the results of the estimation of equation

\[
    pm_{i_1j_1} = \alpha (distance)^{\beta} (size)^{\gamma}
\]  

(70)

<table>
<thead>
<tr>
<th>( )</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficients</td>
<td>1988.19***</td>
<td>-1.57***</td>
<td>-0.20***</td>
</tr>
<tr>
<td>Standard errors</td>
<td>(286.51)</td>
<td>(0.03)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Number of observations</td>
<td>2,044</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* \( p<.1; \) ** \( p<.05; \) *** \( p<.01 \)

The analysis of the propensity to match between workers of occupation \( i_2 \) and jobs of occupation \( j_2 \) \( (pm_{i_2j_2}) \) is similar to that made above for provinces \( (pm_{i_1j_1}) \). Beginning with the case of matches between workers and jobs from the same occupation (contribution group) \( (pm_{i_2j_2}, \text{with no occupational mobility}, \ i_2 = j_2) \), on the left side of Fig. 3 we represent the values of \( pm_{i_2j_2} \) and the size of the corresponding cell for the 10 occupations.

\(^{13}\) The distances to the five extra peninsular provinces are omitted.

\(^{14}\) In our data, \( b \) in equation (68) is not constant, but slightly increasing with size (with no geographical mobility, \( i_1 = j_1 \)). This allows \( pm_{i_1j_1} \) to be decreasing with size (with geographical mobility, \( i_1 \neq j_1 \)). Discrete low values of frequencies also contribute to the inverse relation between propensity to match and size (see footnotes 15 and 16).
Again, the propensity to match within the same occupation is high (clearly above 1; median = 14) for all the occupations, reflecting low occupational mobility, and there is a clear inverse relation between propensity to match and size. As before, the explanation is based on that the high proportion of matches without mobility does not vary much across occupations (75% on average).

The propensity to match between workers and jobs from different occupations ($p_{m_{i_2j_2}}$, with occupational mobility, $i_2 \neq j_2$) is much lower, with all the values lower than 1; median = 0.3. As one would expect, propensities are higher between occupations with similar level of qualification as, for instance, ‘Managers and workers with university degree’ and ‘Technical engineers and qualified assistants’. The size of the cell affects the propensity, but the influence is small as can be observed on the right side of Fig. 315.

We complete the analysis of partial propensities with that of workers to combine location $i_1$ with occupation $i_2$ ($p_{m_{i_1i_2}}$) and of jobs to combine location $j_1$ with occupation $j_2$ ($p_{m_{j_1j_2}}$). As mentioned before, the two variables are highly positively correlated, reflecting the geographical and occupational adjustment between both sides of the market. We show in Table 5 the highest values of $p_{m_{i_1i_2}}$ and $p_{m_{j_1j_2}}$. It should be noted that they seem to concentrate on certain regions (Autonomous Communities), as Galicia, and occupations, as ‘Clerical and workshop heads’.

---

15 The explanation is similar to the case of partial propensity to match between different provinces (see footnote 13).
Table 5. Propensity to combine province with occupation (for workers and for jobs).
(Provinces and occupations corresponding to the highest values)

<table>
<thead>
<tr>
<th>Province</th>
<th>Occupation</th>
<th>$pm_{i_{1/2}}$</th>
<th>$pm_{j_{1/2}}$</th>
<th>Geometric mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coruña, A</td>
<td>Clerical and workshop heads</td>
<td>3.99</td>
<td>3.65</td>
<td>3.81</td>
</tr>
<tr>
<td>Pontevedra</td>
<td>Clerical and workshop heads</td>
<td>3.80</td>
<td>3.61</td>
<td>3.70</td>
</tr>
<tr>
<td>Lugo</td>
<td>Clerical and workshop heads</td>
<td>2.62</td>
<td>4.75</td>
<td>3.53</td>
</tr>
<tr>
<td>Girona</td>
<td>Clerical and workshop heads</td>
<td>2.70</td>
<td>3.96</td>
<td>3.26</td>
</tr>
<tr>
<td>Melilla</td>
<td>Subordinates</td>
<td>3.09</td>
<td>3.33</td>
<td>3.20</td>
</tr>
<tr>
<td>Ourense</td>
<td>Subordinates</td>
<td>2.80</td>
<td>3.13</td>
<td>2.96</td>
</tr>
<tr>
<td>Ourense</td>
<td>Technical engineers and qualified assistants</td>
<td>2.79</td>
<td>2.74</td>
<td>2.77</td>
</tr>
<tr>
<td>Lugo</td>
<td>Technical engineers and qualified assistants</td>
<td>2.63</td>
<td>2.26</td>
<td>2.44</td>
</tr>
<tr>
<td>Ourense</td>
<td>Clerical and workshop heads</td>
<td>1.87</td>
<td>2.71</td>
<td>2.25</td>
</tr>
<tr>
<td>Jaén</td>
<td>Labourers and related trades</td>
<td>2.07</td>
<td>2.23</td>
<td>2.15</td>
</tr>
<tr>
<td>Gipuzkoa</td>
<td>Subordinates</td>
<td>2.14</td>
<td>2.07</td>
<td>2.10</td>
</tr>
<tr>
<td>Lugo</td>
<td>Subordinates</td>
<td>2.25</td>
<td>1.95</td>
<td>2.09</td>
</tr>
<tr>
<td>Bizkaia</td>
<td>Subordinates</td>
<td>2.01</td>
<td>2.17</td>
<td>2.09</td>
</tr>
<tr>
<td>Segovia</td>
<td>Technical engineers and qualified assistants</td>
<td>1.86</td>
<td>2.14</td>
<td>1.99</td>
</tr>
<tr>
<td>Ceuta</td>
<td>Managers and workers with university degree</td>
<td>2.02</td>
<td>1.67</td>
<td>1.84</td>
</tr>
<tr>
<td>Tarragona</td>
<td>Clerical and workshop heads</td>
<td>1.50</td>
<td>2.02</td>
<td>1.74</td>
</tr>
</tbody>
</table>

Now, we analyse the total propensity to match as a function of the partial propensities, on the basis of equation (66) and the corresponding estimation in Table 3. In Table 6, for each type of mobility, we show the maximum, minimum and median values of the total propensity to match. As stated above for partial propensities, total propensity also decreases sharply when mobility increases. We also show the standard deviation of the base-10 logarithm of the total propensity and of the four factors in the right-hand side of equation (66). The standard deviation of the factor logs approximates their contribution to the variation of the log of the total propensity to match. The contribution to the variation of the total propensity to match of $p_{m_{i_{1/2}}}^{\alpha_1}$ (match between locations) is higher than that of $p_{m_{i_{2/2}}}^{\alpha_2}$ (match between occupations), and both are larger than that of $(p_{m_{i_{1/2}}} p_{m_{j_{1/2}}})^\beta$ (propensity of workers and jobs to combine provinces with occupations). The residual factor increases with mobility, but this is mainly due to the lower values of the observed frequencies, which makes the estimation of equation (66) less accurate.
Consider now in more detail the different cases of mobility. We have previously found that, in the case of no mobility ($i_1 = j_1$; $i_2 = j_2$), the partial propensities appearing in the first two factors of equation (66) ($pm_{i_1j_1}^{\alpha_1}$ and $pm_{i_2j_2}^{\alpha_2}$) show a clear inverse relation between propensity to match and size. As a consequence, the total propensity to match draws also an analogous inverse relation, as can be observed on the left side of Fig. 4, with provinces-occupations of bigger size (Madrid-'Labourers and related trades', for example, marked with $\Delta$ in the figure) down to the right and vice versa. The aforementioned high values (well above 1) of the two factors explain the high value of the total propensity to match, as reflected in Table 6 (median = 604).

<table>
<thead>
<tr>
<th>Mobility</th>
<th>Mobility Max</th>
<th>Mobility Min</th>
<th>Mobility Median</th>
<th>Mobility Std. Dev.</th>
<th>$pm_{i_1j_1}^{\alpha_1}$</th>
<th>$pm_{i_2j_2}^{\alpha_2}$</th>
<th>$(pm_{i_1j_1}^{\alpha_1} pm_{i_2j_2}^{\alpha_2})^\beta$</th>
<th>rf</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>41,449</td>
<td>12</td>
<td>604</td>
<td>0.56</td>
<td>0.44</td>
<td>0.33</td>
<td>0.22</td>
<td>0.09</td>
</tr>
<tr>
<td>Occupational (only)</td>
<td>1,281</td>
<td>0.13</td>
<td>13</td>
<td>0.62</td>
<td>0.43</td>
<td>0.38</td>
<td>0.05</td>
<td>0.22</td>
</tr>
<tr>
<td>Geographical (only)</td>
<td>1,859</td>
<td>0.0035</td>
<td>0.76</td>
<td>0.87</td>
<td>0.67</td>
<td>0.33</td>
<td>0.15</td>
<td>0.50</td>
</tr>
<tr>
<td>Geographical and Occupational</td>
<td>404</td>
<td>0.0020</td>
<td>0.18</td>
<td>0.67</td>
<td>0.69</td>
<td>0.30</td>
<td>0.14</td>
<td>0.62</td>
</tr>
</tbody>
</table>

**Table 6. Total propensity to match and factor decomposition.**

Figure 4. Total propensity to match within and between provinces and occupations.\textsuperscript{16}

\textsuperscript{16} The decreasing parallel lines that appear in the right side correspond to discrete low values of $n_{i_1i_2j_1j_2} = 1, 2, 3$, etc.

From (52), (53) and (54) and as size = $p_{i_1}p_{i_2}p_{j_1}p_{j_2}$ we get
To the effect of these two factors, it must be added the effect of \((pm_{i_1i_2}pm_{j_1j_2})^\beta\) and \(rf\) in equation (66), which makes the scatter plot in Fig. 4 more dispersed than on the left side of Fig. 2 and 4. Of these, as shown in Table 6 in this case of no mobility, the most important contribution to the variation of the total propensity to match is that of \((pm_{i_1i_2}pm_{j_1j_2})^\beta\). On the left panel of Fig. 4 we highlight (connected with a line) ten points of highest total propensity for a given size, and all them correspond to those in Table 5 (highest values of \(pm_{i_1i_2}\) and \(pm_{j_1j_2}\)) or are close to these values.

The opposite case to no mobility is that of geographical and occupational mobility \((i_1 \neq j_1; i_2 \neq j_2)\). Again, the inverse relation between propensity to match and size for the first two factors of equation (66) \((pm_{i_1j_1}^{\alpha_1} \text{ and } pm_{i_2j_2}^{\alpha_2})\) implies that the total propensity to match shows also an analogous inverse relation, as can be detected in the right panel of Fig. 4. The low values of the two factors explain the low value of the total propensity to match, with a median \(= 0.18\) well below 1. In Fig. 4 there are highlighted (connected with a line) 22 points of highest total propensity for a given size and the corresponding provinces and occupations are detailed in Table 7 (except the last three rows which we will comment later).

\[
pm_{i_1i_2j_1j_2} = \frac{n_{i_1i_2j_1j_2}}{n_{\text{size}}}
\]

and taking logs

\[
\log pm_{i_1i_2j_1j_2} = \log(\frac{n_{i_1i_2j_1j_2}}{n_{\text{size}}}) - \log \text{size}
\]

\(^{17}\) The decreasing parallel lines corresponding to discrete low values of \(n_{i_1i_2j_1j_2}\) also contribute to the inverse relation between propensity to match and size.
Table 7. Propensity to match with geographical and occupational mobility.
(Provinces and occupations corresponding to highest total propensity for a given size)

<table>
<thead>
<tr>
<th>Worker province</th>
<th>Job province</th>
<th>Distance (km)</th>
<th>Worker occupation</th>
<th>Job occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valencia</td>
<td>Castellón</td>
<td>65</td>
<td>Other clerical workers</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td>Valencia</td>
<td>Castellón</td>
<td>65</td>
<td>3rd class officials and specialists</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td>Álava</td>
<td>Bizkaia</td>
<td>66</td>
<td>Administrative officials</td>
<td>Clerical and workshop heads</td>
</tr>
<tr>
<td>Toledo</td>
<td>Madrid</td>
<td>71</td>
<td>Labourers and related trades</td>
<td>1st and 2nd class officials</td>
</tr>
<tr>
<td>Toledo</td>
<td>Madrid</td>
<td>71</td>
<td>Labourers and related trades</td>
<td>3rd class officials and specialists</td>
</tr>
<tr>
<td>Toledo</td>
<td>Madrid</td>
<td>71</td>
<td>1st and 2nd class officials</td>
<td>3rd class officials and specialists</td>
</tr>
<tr>
<td>Toledo</td>
<td>Madrid</td>
<td>71</td>
<td>Administrative officials</td>
<td>Other clerical workers</td>
</tr>
<tr>
<td>Alicante</td>
<td>Murcia</td>
<td>75</td>
<td>Other clerical workers</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td>Lleida</td>
<td>Tarragona</td>
<td>91</td>
<td>Managers and workers with university degree</td>
<td>Clerical and workshop heads</td>
</tr>
<tr>
<td>Barcelona</td>
<td>Girona</td>
<td>100</td>
<td>Technical engineers and qualified assistants</td>
<td>Clerical and workshop heads</td>
</tr>
<tr>
<td>Gipuzkoa</td>
<td>Bizkaia</td>
<td>119</td>
<td>Clerical and workshop heads</td>
<td>Administrative officials</td>
</tr>
<tr>
<td>Gipuzkoa</td>
<td>Bizkaia</td>
<td>119</td>
<td>Administrative officials</td>
<td>Clerical and workshop heads</td>
</tr>
<tr>
<td>Pontevedra</td>
<td>Coruña, A</td>
<td>121</td>
<td>3rd class officials and specialists</td>
<td>1st and 2nd class officials</td>
</tr>
<tr>
<td>Ourense</td>
<td>Coruña, A</td>
<td>175</td>
<td>Administrative officials</td>
<td>Assistants</td>
</tr>
<tr>
<td>Sevilla</td>
<td>Madrid</td>
<td>538</td>
<td>Other clerical workers</td>
<td>1st and 2nd class officials</td>
</tr>
<tr>
<td>Huelva</td>
<td>Segovia</td>
<td>619</td>
<td>Other clerical workers</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td>Barcelona</td>
<td>Madrid</td>
<td>621</td>
<td>Other clerical workers</td>
<td>Administrative officials</td>
</tr>
<tr>
<td>Barcelona</td>
<td>Madrid</td>
<td>621</td>
<td>Labourers and related trades</td>
<td>Other clerical workers</td>
</tr>
<tr>
<td>Barcelona</td>
<td>Madrid</td>
<td>621</td>
<td>1st and 2nd class officials</td>
<td>Other clerical workers</td>
</tr>
<tr>
<td>Barcelona</td>
<td>Madrid</td>
<td>621</td>
<td>1st and 2nd class officials</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td>Barcelona</td>
<td>Madrid</td>
<td>621</td>
<td>3rd class officials and specialists</td>
<td>Other clerical workers</td>
</tr>
<tr>
<td>Madrid</td>
<td>Barcelona</td>
<td>621</td>
<td>Other clerical workers</td>
<td>Administrative officials</td>
</tr>
<tr>
<td>Segovia</td>
<td>Barcelona</td>
<td>650</td>
<td>1st and 2nd class officials</td>
<td>Administrative officials</td>
</tr>
<tr>
<td>Barcelona</td>
<td>Málaga</td>
<td>997</td>
<td>Clerical and workshop heads</td>
<td>Other clerical workers</td>
</tr>
<tr>
<td>León</td>
<td>Palmas, Las</td>
<td>–</td>
<td>Labourers and related trades</td>
<td>Other clerical workers</td>
</tr>
</tbody>
</table>

All the four factors in equation (66) and Table 6 contribute to those 22 highest values:

a) The kilometric distances between most of the pairs of provinces are short, which makes higher the propensity to match between provinces \( pm_{i,j1} \). In other cases, such as Barcelona-Madrid and vice versa or Sevilla-Madrid, the kilometric distance is longer, but \( pm_{i,j1} \) is high given the distance and the size of the cell, which can be attributed to a greater connection between large cities. Most striking is the case of Huelva-Segovia: minor cities, high distance, but high \( pm_{i,j1} \) (furthermore, concentrated in moving from ‘Other clerical workers’ to ‘Labourers and related trades’).
b) Most of the pairs of occupations have similar levels of qualification (‘Other clerical workers’, ‘Administrative officials’, ‘Clerical and workshop heads’, etc.), which makes higher the propensity to match between occupations \( pm_{ij} \).

c) The propensities of workers and jobs to combine location with occupation are higher on average in these 22 cases than in the rest of pairs province-occupation (higher propensities to combine \( pm_{i_1j_2} \) and \( pm_{j_1i_2} \)).

d) Finally, the residual factor is also higher on average in these 22 cases than in the rest of cases with geographical and occupational mobility.

The residual factor measures which part of the total propensity is not explained by the partial propensities and therefore can be attributed to a specific matching between workers of combined categories \( i_1j_2 \) and jobs of combined categories \( j_1j_2 \). From the analysis of the cases with highest values of \( rf \) it appears that:

a) Mainly, they correspond to shorter distances than in the generality of cases and therefore \( pm_{i_1j_1} \) is also high, which contributes to make the total propensity higher (for instance, in Table 7, Gipuzkoa-‘Administrative officials’ with Bizkaia-‘Clerical and workshop heads’).

b) Occupations such as ‘Administrative officials’, ‘Clerical and workshop heads’ and ‘Other clerical workers’ are prevalent. Two cases reappear permuting the occupations of worker and job for the same provinces.\(^{18}\)

c) In the cases with longer distances the most frequent job province is Barcelona. Also, the residual factor is generally the prominent factor making higher the total propensity (‘pure specific matches’ as, for example, the last three rows in Table 7).

d) As already mentioned above, the observed frequencies with geographical and occupational mobility are low, which reduces the accuracy of results and may contribute to a high \( rf \).

Geographical mobility (only) and occupational mobility (only) are intermediate cases between no mobility and geographical and occupational mobility, and are symmetrical in some respects. In both cases, again, the inverse relation between propensity to match and size for the first two factors of equation (66) implies that the total propensity to match shows an analogous inverse relation, similar to that on both sides of Fig. 4. The values of one of the factors (\( pm_{i_2j_2} \) for geographical mobility, \( pm_{i_1j_1} \) for occupational mobility) are high when

---

\(^{18}\) For example, ‘Clerical and workshop heads’ workers from Gipuzkoa match with ‘Administrative officials’ jobs from Bizkaia, but also ‘Administrative officials’ workers from Gipuzkoa match with ‘Clerical and workshop heads’ jobs from Bizkaia.
the values of the other are low, which results in values of the total propensity in-between those of no mobility and geographical and occupational mobility, as reflected in Table 6. Table 8 presents the provinces and occupations corresponding to the 49 points of highest total propensity for a given size (29 cases of geographical mobility and 20 of occupational mobility).

Let us examine how the four factors in equation (66) and Table 6 contribute to those highest total propensities:

a) In the cases of geographical mobility (only), the kilometric distances between most of the pairs of provinces are short, which makes higher the propensity to match between provinces $pm_{i_1j_1}$. In other cases (Barcelona-Madrid and vice versa) the distance is longer, but $pm_{i_1j_1}$ is high given the distance and the size of the cell, which can be attributed again to a greater connection between big cities. More striking are the cases of Granada-Soria and Cantabria-Granada: small/medium cities, high distance, relatively high values of $pm_{i_1j_1}$, and matches concentrated in ‘Managers and workers with university degree’ and ‘Assistants’ respectively.

In a similar way, in the cases of occupational mobility (only) most of the pairs of occupations have similar levels of qualification, which makes higher the propensity to match between occupations $pm_{i_2j_2}$. Occupations of low level of qualification are prevalent, which can be related to the fact that they have lesser geographical mobility (in this case, total propensity tends to be higher when there is no geographical mobility).
Table 8. Propensity to match with geographical mobility or with occupational mobility.
(Provinces and occupations corresponding to highest total propensity for a given size)

<table>
<thead>
<tr>
<th>Mobility</th>
<th>Worker province</th>
<th>Job province</th>
<th>Distance (km)</th>
<th>Worker occupation</th>
<th>Job occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographical</td>
<td>Valencia</td>
<td>Castellón</td>
<td>65</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td>only</td>
<td>Bizkaia</td>
<td>Álava</td>
<td>66</td>
<td>Technical engineers and qualified assistants</td>
<td>Technical engineers and qualified assistants</td>
</tr>
<tr>
<td></td>
<td>Toledo</td>
<td>Madrid</td>
<td>71</td>
<td>Other clerical workers</td>
<td>Other clerical workers</td>
</tr>
<tr>
<td></td>
<td>Alcante</td>
<td>Murcia</td>
<td>75</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td></td>
<td>Murcia</td>
<td>Alcázar</td>
<td>75</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td></td>
<td>Riaza, La</td>
<td>Álava</td>
<td>86</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td></td>
<td>Riaza, La</td>
<td>Navarra</td>
<td>88</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td></td>
<td>Lugo</td>
<td>Ourense</td>
<td>95</td>
<td>Subordinates</td>
<td>Subordinates</td>
</tr>
<tr>
<td></td>
<td>Ourense</td>
<td>Lugo</td>
<td>95</td>
<td>Clerical and workshop heads</td>
<td>Clerical and workshop heads</td>
</tr>
<tr>
<td></td>
<td>Coruña, A</td>
<td>Lugo</td>
<td>98</td>
<td>Clerical and workshop heads</td>
<td>Clerical and workshop heads</td>
</tr>
<tr>
<td></td>
<td>Barcelona</td>
<td>Girona</td>
<td>108</td>
<td>Clerical and workshop heads</td>
<td>Clerical and workshop heads</td>
</tr>
<tr>
<td></td>
<td>Valladolid</td>
<td>Segovia</td>
<td>111</td>
<td>Managers and workers with university degree</td>
<td>Managers and workers with university degree</td>
</tr>
<tr>
<td></td>
<td>Lleida</td>
<td>Burgos</td>
<td>118</td>
<td>Assistants</td>
<td>Assistants</td>
</tr>
<tr>
<td></td>
<td>Coruña, A</td>
<td>Pontevedra</td>
<td>121</td>
<td>Managers and workers with university degree</td>
<td>Managers and workers with university degree</td>
</tr>
<tr>
<td></td>
<td>Pontevedra</td>
<td>Coruña, A</td>
<td>121</td>
<td>Technical engineers and qualified assistants</td>
<td>Technical engineers and qualified assistants</td>
</tr>
<tr>
<td></td>
<td>Cádiz</td>
<td>Sevilla</td>
<td>125</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td></td>
<td>León</td>
<td>Valladolid</td>
<td>134</td>
<td>Assistants</td>
<td>Assistants</td>
</tr>
<tr>
<td></td>
<td>Córdoba</td>
<td>Sevilla</td>
<td>138</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td></td>
<td>Teruel</td>
<td>Valencia</td>
<td>145</td>
<td>Technical engineers and qualified assistants</td>
<td>Technical engineers and qualified assistants</td>
</tr>
<tr>
<td></td>
<td>Tarragona</td>
<td>Girona</td>
<td>198</td>
<td>Clerical and workshop heads</td>
<td>Clerical and workshop heads</td>
</tr>
<tr>
<td></td>
<td>Sevilla</td>
<td>Málaga</td>
<td>219</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td></td>
<td>Madrid</td>
<td>Barcelona</td>
<td>621</td>
<td>Managers and workers with university degree</td>
<td>Managers and workers with university degree</td>
</tr>
<tr>
<td></td>
<td>Madrid</td>
<td>Barcelona</td>
<td>621</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td></td>
<td>Madrid</td>
<td>Barcelona</td>
<td>621</td>
<td>1st and 2nd class officials</td>
<td>1st and 2nd class officials</td>
</tr>
<tr>
<td></td>
<td>Barcelona</td>
<td>Madrid</td>
<td>621</td>
<td>Labourers and related trades</td>
<td>Labourers and related trades</td>
</tr>
<tr>
<td></td>
<td>Barcelona</td>
<td>Madrid</td>
<td>621</td>
<td>Other clerical workers</td>
<td>Other clerical workers</td>
</tr>
<tr>
<td></td>
<td>Granada</td>
<td>Soria</td>
<td>665</td>
<td>Managers and workers with university degree</td>
<td>Managers and workers with university degree</td>
</tr>
<tr>
<td></td>
<td>Cantabria</td>
<td>Granada</td>
<td>827</td>
<td>Assistants</td>
<td>Assistants</td>
</tr>
<tr>
<td>Occupational</td>
<td>Almeria</td>
<td>Almería</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>only</td>
<td>Cáceres</td>
<td>Cáceres</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ciudad Real</td>
<td>Ciudad Real</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cuenca</td>
<td>Cuenca</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Huelva</td>
<td>Huelva</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jaén</td>
<td>Jaén</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Murcia</td>
<td>Murcia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sevilla</td>
<td>Sevilla</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Barcelona</td>
<td>Barcelona</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cuenca</td>
<td>Cuenca</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cuenca</td>
<td>Cuenca</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cuenca</td>
<td>Cuenca</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molina</td>
<td>Molina</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molina</td>
<td>Molina</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Soria</td>
<td>Soria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Valencia</td>
<td>Valencia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molina</td>
<td>Molina</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molina</td>
<td>Molina</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molina</td>
<td>Molina</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Barcelona</td>
<td>Barcelona</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>León</td>
<td>León</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ávila</td>
<td>Ávila</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b) $pm_{i2j2}$ in the cases of geographical mobility (only) and $pm_{i1j1}$ in the cases of occupational mobility (only) show little variation for a given size (see the left side of Fig. 2 and 3) and, therefore, they are not relevant in their contribution to those highest values.
c) $pm_{i_1i_2}$ and $pm_{j_1j_2}$ are higher on average in these 49 cases than in the rest of pairs province-occupation.

d) Finally, the residual factor is also higher on average in these 49 cases than in the rest of cases with geographical mobility or with occupational mobility.

As mentioned before, the residual factor can be attributed to a specific matching between workers of combined categories $i_1j_2$ and jobs of combined categories $j_1j_2$. From the analysis of the cases with highest values of $rf$ it appears that:

a) Almost completely, they correspond to cases of geographical mobility.

b) Mainly, they correspond to shorter distances than in the generality of cases and therefore $pm_{i_1j_1}$ is also high, which contributes to make the total propensity higher (for example, León-‘Assistants’ with Valladolid-‘Assistants’).

c) Occupations of high level of qualification are prevalent, which can be related to the fact that they have greater geographical mobility.

d) In the cases with longer distances, the ‘residual factor’ is generally the prominent factor making the total propensity higher (‘pure specific matches’ as, for example, the last two rows of the case of only geographical mobility in Table 8 and also ‘Clerical and workshop heads’ from Barcelona to Jaén and from Almería to Gipuzkoa, and ‘Assistants’ from Alicante to Huelva –these last three cases are not shown in Table 8).

e) As already mentioned above, the low observed frequencies reduce the accuracy of results and may contribute to the existence of high values of $rf$.

3.3. Clustering

In our application, the similarity between each pair of categories $i_A-i_B$ (workers of province $i_{iA}$ and occupation $i_{2A}$ and workers of province $i_{iB}$ and occupation $i_{2B}$) is the overlapping or percentage of coincidence of their row profiles or, according to equation (41)

\[
sim_{i_A-i_B} = \sum_{j_1j_2} p_{j_1j_2} \min(pm'_{i_1d_1i_2j_2}, pm'_{i_1i_2d_2j_2})
\]  

(71)

Its value will be between one (if the row profiles are identical) and zero (if their intersection is null).
In an analogous way, the similarity between each pair of categories \( j_A-j_B \) (jobs of province \( j_{1A} \) and occupation \( j_{2A} \) and jobs of province \( j_{1B} \) and occupation \( j_{2B} \)) is the overlapping or percentage of coincidence of their column profiles or, according to equation (44)

\[
\text{sim}_{j_A-j_B} = \sum_{l_1,l_2} p_{l_1,l_2} \min \left( \text{pm}'_{i_1,l_2,1A,l_2A}, \text{pm}'_{i_2,l_1,1B,l_2B} \right)
\]  

(72)

From (71) and (72) we can infer that similarity will be high between categories \( i_A-i_B \) (or \( j_A-j_B \)) with similar values of \( \text{pm}'_{i_1,l_2,1A,l_2A} \) and \( \text{pm}'_{i_2,l_1,1B,l_2B} \) (or \( \text{pm}'_{i_1,l_2,1A,l_2A} \) and \( \text{pm}'_{i_1,l_2,1B,l_2B} \)), which, in turn, can be derived from similar factors in (67). We have commented previously that propensities to match \( \text{pm}_{i_1,l_1} \) and \( \text{pm}_{i_2,l_2} \) are the main factors in this equation. Therefore, \( \text{sim}_{i_A-i_B} \) will be high if the partial propensities to match with each location of jobs are similar for workers of location \( i_{1A} \) and workers of location \( i_{1B} \), and the partial propensities to match with each occupation of jobs are similar for workers of occupation \( i_{2A} \) and workers of occupation \( i_{2B} \). Moreover, \( \text{sim}_{i_A-i_B} \) will be even higher if the categories of one of the two variables are coincident for \( i_A \) and \( i_B \) (for example the same location combined with two different occupations or vice versa). A similar argument applies to \( \text{sim}_{j_A-j_B} \).

The partial similarities between each pair of workers provinces \( i_{1A}-i_{1B} \), workers occupations \( i_{2A}-i_{2B} \), jobs provinces \( j_{1A}-j_{1B} \) and jobs occupations \( j_{2A}-j_{2B} \) are, according to equations (45) and (46)

\[
\text{sim}_{i_{1A}-i_{1B}} = \sum_{l_1} p_{l_1} \min \left( \text{pm}_{i_{1A},l_1}, \text{pm}_{i_{1B},l_1} \right)
\]  

(73)

\[
\text{sim}_{i_{2A}-i_{2B}} = \sum_{l_2} p_{l_2} \min \left( \text{pm}_{i_{2A},l_2}, \text{pm}_{i_{2B},l_2} \right)
\]  

(74)

\[
\text{sim}_{j_{1A}-j_{1B}} = \sum_{l_1} p_{l_1} \min \left( \text{pm}_{j_{1A},l_1}, \text{pm}_{j_{1B},l_1} \right)
\]  

(75)

\[
\text{sim}_{j_{2A}-j_{2B}} = \sum_{l_2} p_{l_2} \min \left( \text{pm}_{j_{2A},l_2}, \text{pm}_{j_{2B},l_2} \right)
\]  

(76)

The values of partial similarities are also between one (if the row profiles are identical) and zero (if their intersection is null).

As we stated before, the similarities between \( i \) categories (in rows) and \( j \) categories (in columns), combined or partial, are highly positively correlated (corr \( \text{sim}_{i_A-i_B} , \text{sim}_{j_A-j_B} \) = 0.839; corr \( \text{sim}_{i_{1A}-i_{1B}} , \text{sim}_{j_{1A}-j_{1B}} \) = 0.863; corr \( \text{sim}_{i_{2A}-i_{2B}} , \text{sim}_{j_{2A}-j_{2B}} \) = 0.742). As it
is convenient to use the same measure of similarity in the clustering process of rows and columns in order to obtain the same grouping on both sides, we will use the arithmetic mean of the above expressions for rows and columns.

\[
sim_{c_A - c_B} = \frac{\sim_{i_A - i_B} + \sim_{j_A - j_B}}{2}
\]

(77)

\[
sim_{c_{1A} - c_{1B}} = \frac{\sim_{i_{1A} - i_{1B}} + \sim_{j_{1A} - j_{1B}}}{2}
\]

(78)

\[
sim_{c_{2A} - c_{2B}} = \frac{\sim_{i_{2A} - i_{2B}} + \sim_{j_{2A} - j_{2B}}}{2}
\]

(79)

where \(c_A - c_B\) represents the corresponding pair of combined categories in rows \((i_A - i_B)\) and in columns \((j_A - j_B)\), and \(c_{1A} - c_{1B}\) and \(c_{2A} - c_{2B}\) the corresponding pair of categories of each variable (provinces and occupations) in rows \((i_{1A} - i_{1B})\) and \((i_{2A} - i_{2B})\) and in columns \((j_{1A} - j_{1B})\) and \((j_{2A} - j_{2B})\).

We relate the similarities of combined categories province-occupation with the partial similarities of provinces and occupations separately using equation (51)

\[
sim_{c_A - c_B} = \sim_{c_{1A} - c_{1B}}^{a_1} \sim_{c_{2A} - c_{2B}}^{a_2} + \varepsilon
\]

(80)

Table 9 presents the results of the estimation of equation (80). We obtain an acceptable fit in the regression and \(a_1, a_2 \approx 1\).

<table>
<thead>
<tr>
<th></th>
<th>(\alpha_1)</th>
<th>(\alpha_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficients</td>
<td>1.09***</td>
<td>1.05***</td>
</tr>
<tr>
<td>Standard errors</td>
<td>(0.0009)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Number of observations</td>
<td>162,735</td>
<td></td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.78</td>
<td></td>
</tr>
</tbody>
</table>

* \(p<.1\); ** \(p<.05\); *** \(p<.01\)

On the basis of equation (80), we can analyse the partial similarities and their contribution to the similarities of combined categories province-occupation. As the clustering process is based on these measures of similarity, we also find a similar relation between the clusters obtained for combined categories province-occupation and those obtained for provinces and occupations separately.
Table 10. Similarities between provinces, occupations and province-occupation categories.

<table>
<thead>
<tr>
<th>Similarities</th>
<th>Max</th>
<th>Min</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between provinces ( (sim_{c_1A \rightarrow c_1B} ) )</td>
<td>0.376</td>
<td>0.019</td>
<td>0.043</td>
</tr>
<tr>
<td>Between occupations ( (sim_{c_2A \rightarrow c_2B} ) )</td>
<td>0.367</td>
<td>0.069</td>
<td>0.178</td>
</tr>
<tr>
<td>Between combined categories ( (sim_{c_A \rightarrow c_B} ) )</td>
<td>Differences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Occupational (only)</td>
<td>0.474</td>
<td>0.011</td>
<td>0.170</td>
</tr>
<tr>
<td>Geographical (only)</td>
<td>0.769</td>
<td>0.0001</td>
<td>0.025</td>
</tr>
<tr>
<td>Geographical and occupational</td>
<td>0.245</td>
<td>0.00002</td>
<td>0.010</td>
</tr>
</tbody>
</table>

We begin analysing the similarity between different provinces \( (sim_{c_1A \rightarrow c_1B}, A \neq B) \). As shown in Table 10, their median value is lower than for similarities between different occupations\(^{19}\). In Fig. 5 we represent the values of \( sim_{c_1A \rightarrow c_1B} \) vs. the kilometric distance between province capitals for the 52 provinces. The figure shows how the similarity decreases as the distance increases.

![Similarity vs. Kilometric Distance between Provinces](image)

**Figure 5. Similarity vs. Kilometric Distance between Provinces.**

As outlined in subsection 2.3, based on this similarity measure, we can carry out a clustering process which orders the provinces (similar provinces join together) and groups them in a lower number of ‘regions’ (clusters of similar provinces). The dendrogram in Fig. 6 shows the clustering of the 52 provinces and their grouping in 16 ‘regions’ and 5 ‘bigger regions’. Fig. 7 shows the corresponding geographical map. By last, Fig. 8 represents the matching map obtained from the clustering process and shows in the biclusters the propensity to match between each pair of regions; a darker cell represents a higher propensity. We can observe, for

\(^{19}\) Minimum values exclude the cases of similarity = 0.
example, that workers and jobs of Madrid+ (Madrid and the surrounding provinces) show a high propensity to match with jobs and workers of Madrid+, but also with those of the other regions. The upper side of the figure represents the final clustering process of these 16 regions and their grouping in 5 bigger regions.

![Figure 6. Clustering of provinces.](image-url)
As already pointed out (see Table 10), similarities between different occupations \((sim_{cA\rightarrow cB}, A \neq B)\) are higher (in median) than similarities between different provinces. The dendrogram in Fig. 9 shows the clustering of the 10 categories of the variable occupation and its grouping in 4 clusters.
Similarities are higher between low-qualification occupations which merge in one cluster, while occupations with higher level of qualification show low similarities and do not merge until the end of the clustering process.

Now we analyse the similarities between combined categories province-occupation (\(sim_{cA,cB}\)) and their relation with partial similarities (\(sim_{cA,c1B}, sim_{c2A,c2B}\)). As shown in Table 10, in accordance with what has been stated for partial similarities, similarities with geographical differences are lower (in median) than with occupational differences, and with both are even lower, as would be expected from equation (80).

The clustering process of the 520 combined categories province-occupation (52 provinces x 10 occupations) gives rise to a dendrogram too large for its full graphical representation, but we will show some portions of it. Fig. 10 represents the matching map, obtained by the clustering process of the original 520 categories into 16 clusters, and showing in each bicluster the propensity to match between each pair of clusters –the highlighted biclusters Andalucía/Andalucía and Andalucía/Cataluña-Huesca are amplified below in Fig. 12 and 13, respectively. The upper side of the figure represents the final clustering process of these 16 clusters and their grouping in 5 bigger clusters. As similarities with occupational differences are higher than with geographical differences, occupations join before provinces, and when we reach 16 clusters only geographical differences remain, so that clusters are geographical regions.
From the comparison of Fig. 8 with Fig. 10, it can be concluded that these regions are very similar to those obtained just from the clustering of provinces. Fig. 11 details the correspondence.

Figure 10. Matching map of clusters of grouped provinces-occupations.

(A darker cell represents a higher propensity)
Fig. 10 gives us an overview of the whole map, but we can zoom on the map to study some parts in more detail. For instance, Fig. 12 and 13 amplify the biclusters Andalucía/Andalucía and Cataluña-Huesca/Andalucía of Fig. 10 (in the case of Andalucía, the amplified biclusters are too large and the figures do not include clusters of other provinces which are similar to those included of Cádiz and Sevilla.

The bicluster Andalucía/Andalucía (Fig. 12) belongs to the main diagonal and shows therefore the high propensity to match with no mobility between clusters (in Fig. 10). In the amplified bicluster we can clearly distinguish four minor clusters. Two of them are obtained merging the eight Andalusian provinces for each of the occupations ‘Managers and workers with university degree’ and ‘Technical engineers and qualified assistants’. These occupations appeared separately in the clustering of occupations while the Andalusian provinces appeared grouped in the clustering of provinces. The other two clusters are obtained merging the rest of the occupations for each of the provinces of Cádiz and Sevilla (and also for the other Andalusian provinces, although they do not appear in Fig. 12). Looking at the map we detect, for instance, that workers and jobs of Cádiz or Sevilla and of occupations

20 Ceuta and Murcia are also merged, jointly with the eight Andalusian provinces, in the occupational group ‘Managers and workers with university degree’.

41
‘Managers and workers with university degree’ or ‘Technical engineers and qualified assistants’ show, in general, a high propensity to match with jobs and workers of the other provinces and occupations.

The bicluster Cataluña-Huesca/Andalucía (Fig. 13) is out of the diagonal and thus shows the low propensities of matches with mobility. In the amplified bicluster we can clearly distinguish the same four minor clusters on the side of Andalucía and eight of them on the side of Cataluña-Huesca. Of these eight, three are obtained merging the four Catalan provinces for the occupations ‘Managers and workers with university degree’, ‘Technical engineers and qualified assistants’21 and ‘Clerical and workshop heads’, that appeared separately in the clustering of occupations while the Catalan provinces appeared grouped in the clustering of provinces22. The other five are obtained merging the rest of the occupations for the four Catalan provinces and Huesca. Looking at the map we can find, for example, that jobs of Barcelona and of occupations ‘Technical engineers and qualified assistants’ or ‘Managers and workers with university degree’ or ‘Other clerical workers’ exhibit, in general, a high propensity to match with workers of the different provinces and occupations of Andalucía.

---

21 Castellón is also merged, jointly with the four Catalan provinces, in the occupational group ‘Technical engineers and qualified assistants’.
22 With Aragonese provinces.
Figure 12. Matching map of bicluster Andalucía/Andalucía.
(A darker cell represents a higher propensity)
Figure 13. Matching map of bicluster Andalucía/Cataluña-Huesca.
(A darker cell represents a higher propensity)
In general, we detect a clear relation between the clusters obtained for combined categories province-occupation and those obtained for provinces and occupations separately. The specific analysis of the propensity to match in a particular cell requires more detail to take into account all the variables (mobility, size, low frequencies which reduce the accuracy of results, etc.) that we have considered in subsection 3.2.

4. Conclusions

In this paper we develop some previous works by inserting their contents into the more general framework of contingency tables and dealing with the dimensions problem generated by the combination of the multiple characteristics that define each category of workers and jobs. We apply our results to the Spanish labour market, relying on a database of individual microdata of considerable size. The first part of our study is methodological and the second part is the application of the methodology to the Spanish labour market.

After presenting the characteristics of the contingency tables that are our subject of study, we develop the concept of propensity to match between each row category (worker category in our application) and each column category (job category in our application) in two-way contingency tables. Using a ‘Cobb-Douglas’ functional form we estimate the total propensity to match between each row combined category and each column combined category as a multiplicative function of the partial propensities to match between row individual categories and column individual categories. In this way we disentangle the mix of the effects from the multiple interactions between the different variables. In our application we obtain that total propensity between combined categories province-occupation is approximately the product of the partial propensity to match between provinces and the partial propensity to match between occupations (additionally, another factor reflects the propensity of workers and jobs to combine provinces with occupations). Partial and total propensities decrease as the distance between different provinces and between levels of qualification increases (low geographical and occupational mobility). In other cases, where the propensity to match is high given the distance, this fact can be attributed to a greater connection between large cities. Occupations of high (low) level of qualification have greater (lesser) geographical mobility –see Núñez et al. (2014)–, and, in this case, total propensity tends to be higher when there is (there is no) geographical mobility. We detect also some ‘pure specific matches’ where the residual factor is the prominent factor making the total propensity higher. This can be attributed to a specific matching between the corresponding combined categories of workers and jobs, but we must
take into account that the low observed frequencies reduce the accuracy of results and may contribute to a high residual factor.

We apply the biclustering procedures to our contingency table using a measure of similarity that can be linked to the Manhattan or City Block distance metric. Based on this similarity measure, we can carry out a clustering process which orders the categories (in rows and columns), and groups them in a lower number of clusters and the cells in a lower number of biclusters. Biclustering contributes to solve the dimensions problem reducing the number of rows, columns and cells in the contingency table and increasing cell frequencies.

We analyse the partial similarities and, using again a ‘Cobb-Douglas’ functional form, their contribution to the similarities of combined categories of the different variables (province-occupation in our application). As the clustering process is based on these measures of similarity, we also get a similar relation between the clusters obtained for combined categories of the different variables and those obtained for each variable separately.

In our application we find that similarities between combined categories province-occupation are approximately the product of the partial similarity between provinces and the partial similarity between occupations. Similarities between combined categories province-occupation and partial similarities decrease as the distance between different provinces and between levels of qualification increases.

Using the partial similarities between provinces, we carry out a clustering process which orders the 52 provinces (similar provinces join together) and groups them in 16 ‘regions’ (clusters of similar provinces) and, finally, in 5 ‘bigger regions’. Likewise, using the partial similarities between occupations, we carry out a clustering process which orders the 10 categories of the variable occupation and groups them in 4 clusters. Using the similarities between combined categories province-occupation, we carry out a clustering process which orders the 520 combined categories and groups them in 16 clusters and, finally, in 5 bigger clusters. As similarities with occupational differences are higher than with geographical differences, occupations join before provinces, and when we reach 16 clusters only geographical differences remain, so that clusters are geographical areas that are very similar to those obtained just from the clustering of provinces. They are also similar to the official Autonomous Communities, although there are some significant differences (for example, Madrid ‘absorbs’ some provinces of the adjacent Communities).
We represent the contingency table as a ‘matching map’, obtained from the original contingency table by the clustering process and showing in the biclusters the corresponding propensity to match. We use different levels of ‘zoom’ depending on whether you want to have an overview of the whole map or a more detailed view of a particular area. In general, we find a clear relation between the clusters obtained for combined categories province-occupation and those obtained for provinces and occupations separately.

Worker mobility, geographical or occupational, and the availability of enough information are important requirements for effective labour matching, and a prominent part of active labour market policies. The matching maps can help show job seekers and firms where successful matches have taken place. However, its usefulness is directly related to the relevance of the variables considered in the labour matching context. The data from the MCVL that we have used, and other public databases on labour matches, do not include data as relevant as, for example, workers’ academic degrees and firms’ identities (which would allow knowing their main features). Having access to this kind of data, the empirical tools proposed in this paper may be more useful in helping searchers in the labour market.

References


Stockburger (2016) expresses this idea with a simple but illustrative example: ‘For example, a person might wish to predict how an animal would respond to an invitation to go for a walk. He or she could be given information about the size and weight of the animal, top speed, average number of hours spent sleeping per day, and so forth and then combine that information into a prediction of behavior. Alternatively, the person could be told that an animal is either a cat or a dog. The latter information allows a much broader range of behaviors to be predicted. The trick in cluster analysis is to collect information and combine it in ways that allow classification into useful groups, such as dog or cat’. For instance, in our case ‘dogs and cats’ could be workers’ academic degrees and firms’ identities.


Acknowledgments:

We thank financial support from the Andalusian Council of Economics and Knowledge (SEJ-513 PAIDI Research Group).