The Use of Multiple Correspondence Analysis to Explore Associations Between Categories of Qualitative Variables and Cancer Incidence

Dídac Florensa, Pere Godoy, Jordi Mateo, Francesc Solsona, Tere Pedrol, Miquel Mesas, and Ramon Piñol

***Abstract*—Background: Previous works have shown that risk factors for some kinds of cancer depend on people's lifestyle (e.g. rural or urban residence). This article looks into this, seeking relationships between cancer, age group, gender and population in the region of Lleida (Catalonia, Spain) using Multiple Correspondence Analysis (MCA). Methods: The dataset analysed was made up of 3408 cancer episodes between 2012 and 2014, extracted from the Population-based Cancer Registry (PCR) for Lleida province. The cancers studied were colon and rectal (1059 cases), lung (551 cases), urinary bladder (446 cases), prostate (609 cases) and breast (743 cases). The MCA technique was applied and used to search relationships among the main qualitative features. The basic statistics were the percentage explaining (variance), the inertia and the contribution of each qualitative variable. Results: General outcomes showed a low and moderate contribution of living in rural areas to colorectal and male prostate cancer. Males in urban areas were slightly and heavily affected by lung and urinary bladder cancer respectively. The analysis of each cancer provided additional information. Colorectal cancer greatly affected males aged $<$60, urban residents aged 70–79, and rural females aged $\geq$ 80. The impact of lung cancer was high among urban females $<$60, moderate among males aged 70–79 and high among rural females aged $\geq$ 80. The results for urinary bladder cancer results were similar to those for lung cancer. Prostate cancer affected both the $<$60 and $\geq$80 age groups significantly in rural areas. Breast cancer hit the 70–79 group significantly and, somewhat less so, rural females aged $\geq$ 80. Conclusions: MCA was a significant help for detecting the contributions of qualitative variables and the associations between them. MCA has proven to be an effective technique for analyzing the incidence of cancer. The outcomes obtained help to corroborate suspected trends, as well as detecting and stimulating new hypotheses about the risk factors associated with a specific area and cancer. These findings will be helpful for encouraging new studies and prevention campaigns to highlight observed singularities.**

***Index Terms*—Cancer, cancer registry, multiple correspondence analysis, rural, urban.**

I. Background

CANCER is the second leading cause of death globally. Between 30–50% of cancers can currently be prevented by avoiding risk factors and implementing existing evidence-based prevention strategies. The continuous rise of this disease over recent decades is attributed to the impact of aging among an increasingly elderly population [1].

Cancer recording is considered a key factor in controlling the disease [2]. The purpose of the registers is to detect and fully record all cases of cancer diagnosed among the residents of the reference area [2], [3]. There are three population-based cancer registers (PCR) in Catalonia (Spain), these being the PCRs of the provinces of Lleida, Girona and Tarragona [4]. Barcelona is the fourth province. However, it has no PCR. These records indicate the existence of territorial differences that would need to be studied. Recent studies suggest differences in the incidence of cancer, temporal trends, and mortality among urban and rural areas which are attributable to exposure to different risk factors, access to screening programs, and regular diagnosis and treatment [5]. Specifically, the population of the Lleida region presents life styles, risk factors and work activity which can be traduced to a specific incidence for certain types of cancer. Nearly half of the population of Lleida province live in rural areas. As a consequence, their lifestyle is different from that of the more urban populations in other Catalan provinces. A peculiarity of this region is the work environment. In rural areas,

Manuscript received June 29, 2020; revised November 4, 2020, December 17, 2020, and March 17, 2021; accepted April 11, 2021. Date of publication April 15, 2021; date of current version September 3, 2021. This work was supported in part by the Industrial Doctorate Program of the Government of Catalonia under Contract 2019-DI-43 in part by the Ministerio de Economía y Competitividad under Contract TIN2017-84553-C2-2-R, and in part by the Generalitat de Catalunya (some of the authors are members of the research group 2014-SGR163). *(Corresponding author: Didac Florensa.)*

Dídac Florensa, Jordi Mateo, and Francesc Solsona are with the Department of Computer Science, University of Lleida, 25001 Lleida, Spain (e-mail: didac.florensa@gencat.cat; jordi.mateo@udl.cat; francesc.solsona@udl.cat).

Pere Godoy is with Epidemiology Service, Department of Health, 25006 Lleida, Spain, and also with CIBER Epidemiology and Public Health (CIBERESP), 25006 Lleida, Spain (e-mail: pere.godoy@gencat.cat).

Tere Pedrol is with the Health Department, Population-Based Cancer Registry in Lleida, 25006 Lleida, Spain (e-mail: mtpedrol.ics.lleida@gencat.cat).

Miquel Mesas is with the Department Computer, Santa Maria University Hospital, 25198 Lleida, Spain (e-mail: mmesas@gss.cat).

Ramon Piñol is with the Department of Health, Catalan Health Service, 08023 Barcelona, Spain (e-mail: rpinol@catsalut.cat).

Digital Object Identifier 10.1109/JBHI.2021.3073605

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

the main activity is the agri-food industry and, in urban areas, it is service sector activities such as education, health and catering.

In the literature, there are several reports that present the incidence of cancer in rural and urban areas. In 1992, the University of North Carolina presented a rural-urban pattern study of cancer mortality [6] and explained differences in its incidence among rural versus urban populations. It concluded that cancer is diagnosed at more advanced and more disseminated stages of the disease in rural populations because they are typically older, less educated, poorer and have less access to such health care services as early-cancer detection. Potential explanations were given for lower overall incidence rates in rural areas compared with urban zones. These include smoking (more prevalent in urban areas) and exposure to environmental pollutants. Whitney E Zahnd *et al* ([5]) presented a report about rural-urban differences in cancer incidence and trends in the United States. The study analyzed age-adjusted incidence rates, ratios and annual percentage change (APC) for all cancers detected between 2009 and 2013. Concretely, this report concludes that cancer rates associated with modifiable risks-tobacco, human papillomavirus, and some preventive screening modalities (e.g., colorectal and cervical cancers)- were higher in rural settings compared with urban populations. Next, the work in [7] concluded that, although cigarette smoking is the primary cause of lung cancer, there are other risk factors which may differ by geographic region. These include passive smoking, exposure to indoor radon and asbestos. Finally, [8] present a work investigating urban-rural variations in the incidence of several cancers after adjusting them for socioeconomic status. This interesting article concluded that the risk of some cancers varied with area and gender. For example, the risk of prostate cancer was higher in rural areas and as was that of breast cancer in females in urban areas.

Recently, the PCR team in Lleida presented a descriptive-analytic study highlighting the preliminary results of the impact and incidence of cancer in urban and rural areas [9]. The article compared the number of cancer cases between rural and urban areas according to the crude data rates from the Catalan Population-based Cancer Registry. Tumour ranking and rates obtained in the different areas of the province of Lleida suggested that some cancers have particular features that should be investigated. Jointly with this incipient work, the related literature [5]–[8] has led us to study the incidence of major cancers by rural and urban areas. Many efforts have been made to measure the incidence of cancer by using such traditional methods as density rates, annual rates or the Spearman rank correlation coefficient. However, none of them has provided enough evidence of a relation between cancer and population.

To address these limitations, this paper proposes studying the differences between urban and rural populations. The method proposed is the application of Multiple Correspondence Analysis (MCA) to stimulate new hypotheses and relations between the characteristics of the patients and the incidence of cancer in the province of Lleida.

As the main contribution of this study, we propose the use of MCA as a statistical technique to search for associations between the registered data for cancer in the province of Lleida (Catalonia). This province has a good balance between rural and urban populations and the dataset is mainly made up of categorical variables. In [10], the authors asserted that MCA helped them to classify the degree of tumor regression with a categorical dataset. This led us to assess our challenge with the same statistics using MCA. The outcomes obtained demonstrate the usefulness of this technique in this kind of data analysis, made up exclusively of categorical variables.

II. Methods

The Population-based Cancer Registry (PCR) of the health region of the province of Lleida (HRPLL) was the basis for this descriptive epidemiological study into cancer. The main information sources were hospital records (ICD-9 codes-140.0 to 208.9) and reports from pathological anatomy. Before extracting the information for this study, the cases were reviewed and validated using ASEDAT[1]. Then, an accurate description of the data and basic concepts of the MCA statistical technique used in this work are explained in this section.

A. Data

Lleida is the largest province in Catalonia with a population density of 36 people per square kilometre. More specifically, the population was 438,001 in 2014 [12], 221 891 men and 216,110 women. Approximately half of the population lives in rural areas. In accordance with [13], people living in cities with a population of more than 10 000 are classified as "urban" in this study and the rest as "rural".[2] In 2014, the respective urban and rural populations were 199,300 and 238,701. Thus, this is a well-balanced dataset for studying differences between urban and rural populations in the risk-factors for cancer.

The data are made up of the information registered between 2012-2014 in the Lleida PCR [14], [15] for cancer patients in the main hospitals in the health region of the province of Lleida. These are the Arnau de Vilanova University Hospital (HUAV) and the Santa Maria University Hospital (HUSM). The study is GDPR[3]-compliant, maintaining the anonymity of the patients. Cancer episodes were recorded according to international criteria. These go from the case definition to the operation system and the final results obtained in order to ensure reliability, the validation of the data and comparison with other hospital registers.

The initial dataset consisted of 3,423 new cancer diagnoses in the HRPLL during 2012-2014. After applying data cleaning by using the Box Plot technique to discard statistical outliers, the data collection became 3408 cancer diagnosis (See box plot graphs in the Github repositoy [17]). These box plots graphs are based on each cancer and by age and population and gender. As Figure 1 in the annex shows, this allows outliers for colorectal cancer by age and gender to be detected. And so

[1]ASEDAT: Software of the Catalan Institute of Oncology to select, extract and validate cancer data [11]

[2]The Spanish National Statistics Institute (Spanish initials: INE) has defined rural areas as those with fewer than 2,000 inhabitants; semi-urban areas as those with between 2001 and 10,000 inhabitants; and urban areas as those with more than 10,000 inhabitants.

[3]GDPR: General Data Protection Regulation (EU)

on in the rest of the graphs. This technique uses the median, approximate quartiles, and the lowest and highest data points to convey the level, spread, and symmetry of a distribution of data values [16]. In addition, all the scripts implemented for data cleaning (done with Python) and data analysis (done with R) can be freely downloaded from this Github repository [17]. All the data provided in this link were generated randomly.

Each register contains the following fields: **age group** (<60, 60–69, 70–79, ≥ 80); **gender** (male, female); **population** (rural, urban) and **cancer** type. Only the five most frequent types of cancer were analysed (see incidence tables in the Github repository [17]). These being colon and rectal (1059), urinary bladder (446), breast (743), prostate (609) and lung (551). Gender was divided between males (2088) and females (1319). The population was divided into rural (1821) and urban (1587). Finally, age was divided into 4 balanced intervals: <60 years old with 834 cases, 60 to 69 years old with 927 cases, 70 to 79 with 968 cases and aged ≥ 80 with 679 cases.

B. Statistics

All the information presented was analyzed using Multiple Correspondence Analysis (MCA), an extension of Correspondence Analysis (CA). MCA is an unsupervised learning algorithm for visualizing the patterns in large tables and for multi-dimensional categorical data [18]. This method can be used to describe, explore, summarize and visualize information contained on individuals described by categorical variables within a data table [19]. Unlike CA, MCA can deal with more than one categorical variable. This is the main advantage of the MCA technique. In our case, MCA was firstly used to evaluate the relationships between the four features. It was then used to evaluate the relationships between population, age and gender for each cancer. Associations between features were represented graphically [10]. The graphs aim to visualize the similarities and/or differences in the profiles simultaneously, identifying those dimensions that contain most of the data variability. Features or their categories close to each other are significantly related statistically.

The factors produced were interpreted with the help of various statistical coefficients which complemented each other to provide a better interpretation. The most common and important are inertia, the eigenvalue and the contribution and factorial coordinates. Inertia is a measurement of the dispersion of the set of computed distances between points. Analogously, in Principal Correspondence Analysis (PCA), inertia corresponds to the explained variance of dimensions. The eigenvalue allows the inertia that a specific category produces to be quantified. The contribution enables us to consider how much influence a category has in determining a certain percentage relative to the entire set of the active category. The percentage coordinates (x- and y-axis) of the graph enable the category points in a graph to be represented and established. In MCA, the distance between two or more categories of different variables can be interpreted in terms of the associations and correlations between these. If two categories present high coordinates and are close in space, this means that they tend to be directly associated. If two categories present high coordinates but are distant from each other (e.g. they have opposite signs), this means that they tend to be inversely associated. If two categories present the same coordinate sign, they can be related to each other [20], [21].

Thus, the graphic depiction aims to visualize the similarities and/or differences in the profiles simultaneously, identifying those dimensions that contain most of the data variability. These MCA representations can be read like those from a PCA: the coordinates of a product are its values for the common factors; the coordinates of a variable are its correlation with these factors [22]. Categories depicted in the same direction on the dimension will be significantly related statistically and have patterns of relative frequencies. This association is also valuable statistically when the points are located far from the origin of the graph, representing a mean, uninformative profile [23], [24].

In this study, the MCA method was applied in scripts performed with R [25], an open-source programming language and environment for statistical computing and graphics. It provides a wide variety of statistical and graphical techniques, and is highly extensible. Specifically, the main library used to implement the methods and obtain the results was FactoMineR [26].

III. Results

In this section, we first present a general analysis of the results obtained from applying the MCA technique to the dataset presented in section II. Then, a similar analysis was applied to each cancer to evaluate these in isolation. It is important to clarify the differences between contribution and correlation to understand and interpret the results and figures presented in this section. The contribution is used to denote which variables explain better the variations in the data set and are most important in the construction of the axes. In contrast, correlation represents the relation between two variables or, in other words, the degree of influence of one variable compared with the other.

A. Multiple Correspondence Analysis for All Cancers

The variance obtained was 20.5% (eigenvalue: 0.46) for dimension 1 (x-axis) and 12.7% (eigenvalue: 0.285) for the second one (y-axis). The inertia (sum of the variances) for these two dimensions was 33.2%. Age variance scored 0.259 and 0.582 in dimensions 1 (x-axis) and 2 (y-axis) respectively. Cancer was 0.809 and 0.443, gender 0.765 and 0.007, and population 0.008 and 0.107. The variable that gave the worst results for percentage explanation was population.

Similar results were obtained when discarding the gender variable. The variances in this case were 16.3% (eigenvalue: 0.433) for dimension 1 and 14.1% (eigenvalue: 0.376) for dimension 2, and an inertia for these two dimensions of 30.4%. The percentages of variances explained for population were 0.062 and 0.035.

Removing the age variable, the variances were 28.4% (eigenvalue: 0.568) and 17.3% (eigenvalue: 0.346) for dimensions 1 and 2 respectively. Thus, the inertia for these dimensions was 45.7%. This was the two-dimension combination (the dimensions are ranked with the variance) that gave the highest inertia.

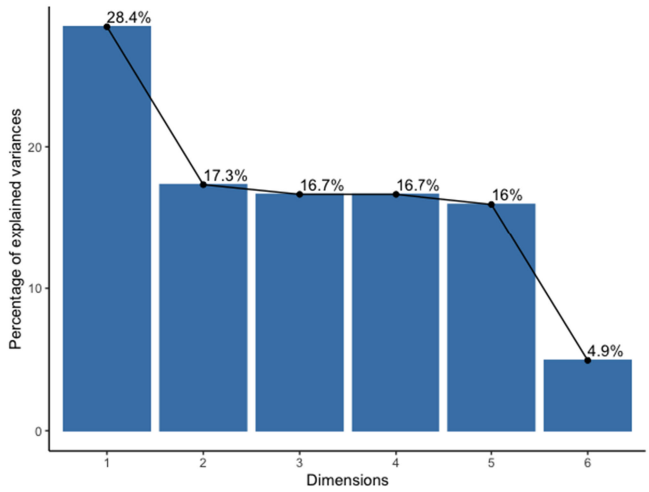

Fig. 1. Percentage of explained variances of the overall dimensions.

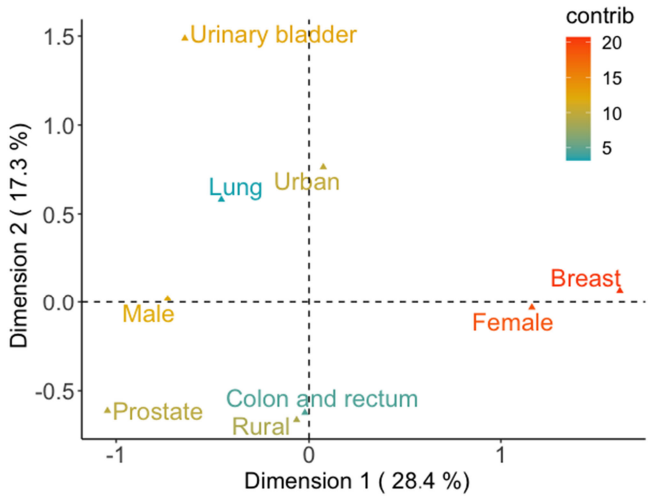

Fig. 2. Two-dimensional MCA plot. Correlations between the variables.

Each variable variance usually increases with the inertia. Fortunately, in this case, the population variances (0.004 and 0.506) improved significantly. This was also the best combination for population, the main goal of the present work.

Fig. 1 shows the variances of the overall dimensions (6) for the combinations of variables obtained. Note that the dimensions are ranked in descending order. It can be seen that dimensions 1 and 2 have variances of 28.4% and 17.3% respectively. The sum of the variances of the overall dimensions is 100%. In this figure, the main idea was to show the percentage of explained variance in every dimension and not the influence of all the variables.

Fig. 2 presents the results of the MCA algorithm in a two-dimensional plot (x- and y-axis representing dimensions 1 and 2 respectively) that shows the correlations between the variables. A two-dimensional plot gives more information about correlations between variables than higher dimensional ones. Thus, no higher dimensional-results were presented. It can be seen that colorectal cancer and rural are very close and appear in the negative y-axis (dimension 2). This means that they are

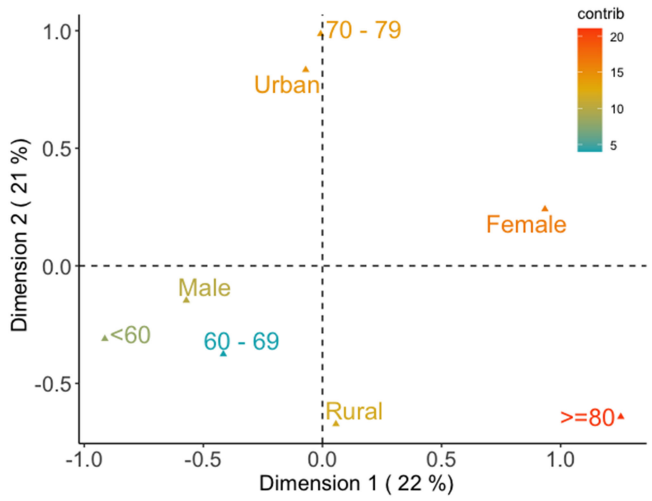

Fig. 3. **Colorectal** cancer. The positive and negative x-axis (representing gender variable) depicts females and males. The positive and negative y-axis (representing population variable) depicts urban and rural.

correlated. Lung and urinary bladder cancers appear on the positive y-axis (dimension 2) where urban contributes and on the negative x-axis (dimension 1) where males appear. This suggests that these cancers are correlated with urban males. Moreover, prostate appears on the negative y-axis (dimension 2) meaning that it is significant in rural areas. Finally, the only breast cancer is correlated with females, with the same contribution in both areas.

The general outcomes showed a low contribution for colorectal cancer in rural areas. The lowest contributions are depicted in blue in the ranking. No differences between gender are observed, due to the location of coordinate 0 on the x-axis. Prostate cancer (its ranked color is located in the middle of the key) affected males in rural areas moderately. A low affection of lung cancer was observed in urban males. Urinary bladder cancer affected urban dwellers severely, mostly males. Finally, as expected, breast cancer heavily affected females independently of the area.

B. Multiple Correspondence Analysis by Cancer

This section presents the MCA results for colorectal, lung, urinary bladder, prostate and breast cancers.

The first cancer studied was **colorectal** (Fig. 3). The variance obtained for the first dimension was 22% (eigenvalue: 0.366) and 21% for the second dimensions (eigenvalue: 0.349). The total inertia was 43%. The correlation between the population variable and dimensions was 0.003 on the first and 0.56 on the second. The gender correlation was 0.533 (dimension 1) and 0.035 (dimension 2), and the age group correlation was 0.561 (dimension 1) and 0.453 (dimension. 2). The urban population was represented on the positive y-axis (29.55% of the total category contributions in dimension 2) and the rural on the negative y-axis (23.85% of the total category contributions in dimension 2). Gender is represented on the x-axis (dimension 1). The female contribution was 30.13% on the positive x-axis and

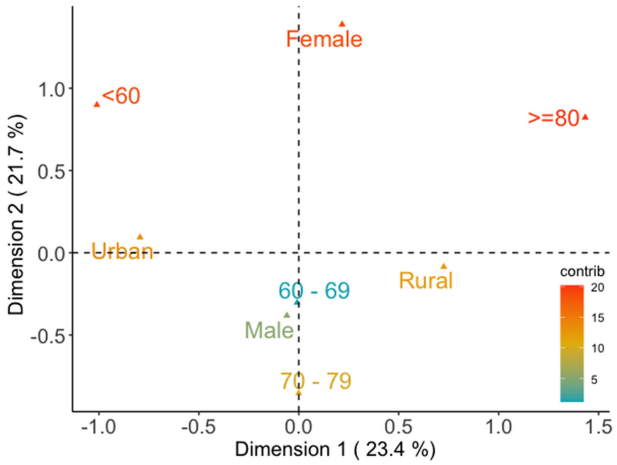

Fig. 4. **Lung cancer**. The positive and negative x-axis (representing population variable) depicts rural and urban respectively. The positive and negative y-axis (representing gender variable) depicts females and males respectively.

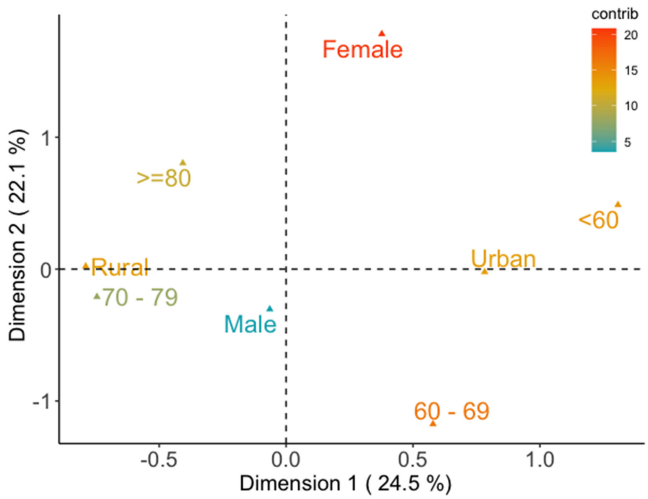

Fig. 5. **Urinary bladder** cancer. The positive and negative x-axis (representing population variable) depicts urban and rural respectively. The positive and negative y-axis (representing gender variable) depicts females and males respectively.

the male contribution was 18.43% on the negative (dimension 1). The group aged ≥ 80 contributed 32.62% in dimension 1 and 8.96% in dimension 2. The contribution of the 70-79 group was 0.001% on the former and 28.85% on the later. The contributions of those aged 60-69 were 4.31% and 3.67%, and the group aged <60 contributed 14.12% and 1.70%.

The 70-79 age group was closely related to the urban population, regardless of gender. This is a very significant result because it has an important y-axis component (close to 1). In the <60 age band, it affected males slightly more. The most important result was the ≥ 80 age group, where there is a higher incidence among women. The incidence among the 60-69 age group was not significant but mainly affected men.

Fig. 4 shows the results obtained for **lung** cancer. The variance for dimension 1 was 23.4% (eigenvalue: 0.390) and 21.7% for dimension 2 (eigenvalue: 0.362), so the total inertia was 45.1%. In this study, the population correlation was 0.576 on dimension 1 and 0.008 on dimension 2, the gender correlation was 0.012 and 0.530, and finally, the age group was 0.581 and 0.548. Regarding the categories variables, urban areas contributed on the negative x-axis (dimension 1) with 25.71% and the rural with 23.48% on the positive x-axis. In the case of gender, the male contribution was 10.53% on the negative y-axis (dimension 2) and the female contribution was 38.24% on the positive y-axis. The ≥ 80 age group contributed 29.28% in the first dimension and 10.37% in the second. The contributions of the 70-79 age group were 0.01% and 20.21%, then 0.019 and 2.51% for the 60-69 age group and the <60 age group contributed 20.40% and 17.38%.

The high position of urban females <60 shows that the contribution of this relation was very high. A moderate significance can be observed for males aged 70-79. Finally, the contribution for rural females aged ≥ 80 reached the same significance as urban females <60.

In **urinary bladder** cancer (Fig. 5), the variance for the first dimension was 24.5% (eigenvalue: 0.407) and 22.1% (eigenvalue: 0.367) for the second, and in consequence, the total inertia was 46.6%. The population correlation was 0.617 on dimension 1 and 0.0004 on dimension 2, the gender correlations were 0.024 and 0.542 and those for the age group were 0.581 and 0.559. The urban category contributed 25.13% on the negative x-axis and the rural on the positive x-axis with 25.36% (dimension 1). The male category contribution was 7.16% on the negative y-axis and that of female was 42.01%. The ≥ 80 age group contributed with 4.12% in dimension 1 and 17.86% in dimension 2, the 70-79 age group contributed 13.74% and 1.2%, the 60-69 age group with 6.13% and 28.08%, and the <60 age group with 23.50% and 3.61%.

The results for urinary bladder cancer were similar to those for lung cancer. Specifically, females aged <60 contributed moderately in urban areas (23.50% in dimension 1 and 3.61% in dimension 2). In the 60-69 age cohort, it affected men in urban areas moderately but this incidence decreased among those aged 70-79 living in rural areas (13.74 in dimension 1 and 1.20% in dimension 2). The contribution of men between 60-69 was 6.13% in dimension 1 and 28.08% in dimension 2. In the ≥ 80 age group, rural women were slightly affected (4.12 in dimension 1 and 17.86% in dimension 2).

Moderate importance was moved to urban males aged 60-69, dropping when reaching the 70-79 age group, in the rural zone. And, attenuated importance was to rural females aged ≥ 80.

Females are ruled out of **prostate** cancer (Fig. 6). The variance for the first dimension was 25% (eigenvalue: 0.5) and 25% (eigenvalue: 0.5) in the second, resulting in total inertia of 50%. The population correlation for dimension 1 was 0.527 and this was 0 for dimension 2, and the correlations for age group were 0.527 and 1. In this case, the gender variable was not included because this type of cancer only affects men. The contribution of the urban category was 28.32% on the positive x-axis and the rural contribution was 21.67% on the negative x-axis (dimension 1). The ≥ 80 age group's contribution was 27.12% in the first dimension and 30.24% in the second. The contributions of the

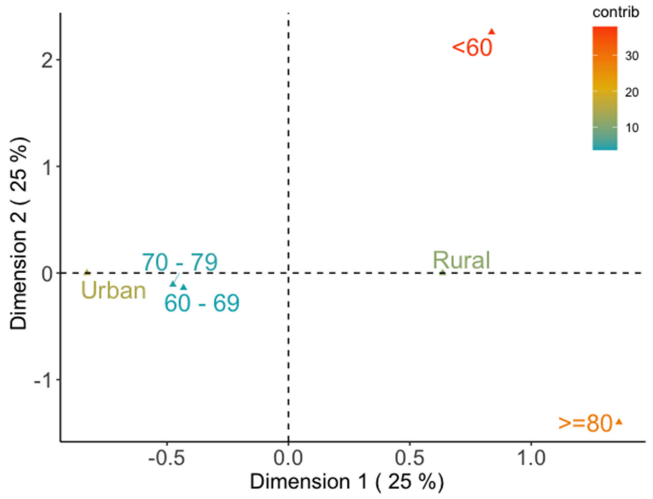

Fig. 6. **Prostate** cancer. The positive and negative x-axis (representing population variable) depicts rural and urban respectively. The y-axis has no meaning on this occasion.

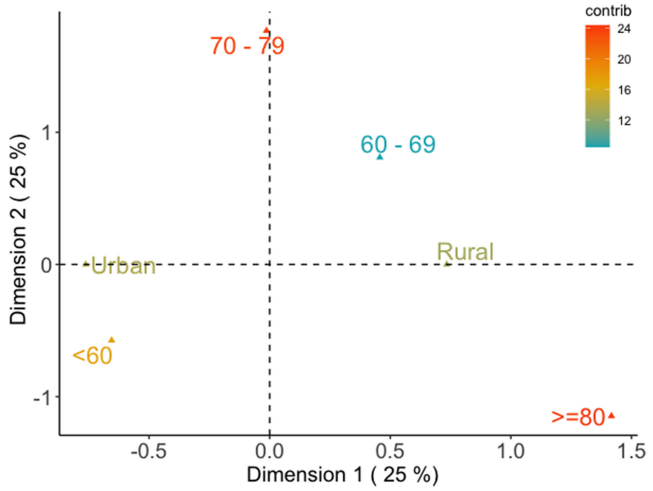

Fig. 7. **Breast** cancer. The positive and negative x-axis (representing population variable) depicts for rural and urban respectively. The y-axis has no meaning on this occasion.

70-79 age group were 7.58% and 0.413%, for 60-69 age group, 6.34% and 0.691%. Finally, the <60 age group contributed 8.94% and 68.64%.

In contrast to the previous cancers analyzed, rural males <60 suffered heavily. Hardly any contributions appeared in the 60-69 and 70-79 urban age ranges. In the absence of females, prostate cancer affected rural males aged ≥80 significantly.

Fig. 7 shows the results obtained when applied the MCA to **breast** cancer for females only. The variance obtained for the first dimension was 25% (eigenvalue: 0.559) and 25% (eigenvalue: 0.5) for the second (total inertia was 50%). The population correlation was 0.559 for dimension 1 and 0.0 for dimension 2, and the age group correlations were 0.559 and 1.0. As explained in subsection II-A, gender was not considered. The contribution of the urban category was 25.47% on the negative x-axis and rural contribution was 24.52% on the positive x-axis (dimension 1). The ≥80 age group's contribution was 27.84% in the first dimension and 20.42% in the second. The 70-79 age group contributed 0.002% and 50.25%. For the 60-69 age group, the percentages were 3.85% and 13.58%, and finally, for the <60 age group, 18.29% and 15.73%.

In the data resulting after applying the screening technique, breast cancer only presented females cases even though males can also suffer from it [27]. This cancer affects rural females aged <60 moderately. Among the 60-69 age group, urban women were hardly affected. The group which contributed the most and with a great significance was those aged 70-79, although the type of population did not influence the results. In the ≥80 group there was, as usual, a high incidence among rural females.

Figs 8 and 9 show the contributions obtained for all categories of cancer in the first and second dimensions, respectively. The figures enable the categories that contribute significantly to be detected and interpreted. They also allow the associations to be detected by the contribution in the same dimension. As they show, the x-axis represents each cancer and the y-axis, the contribution. The categories are represented in each stacked bar. For example, in Fig. 8, colorectal cancer, the ≥80 age group suggests an association between gender because it presents a higher contribution than the others groups. However, in Fig. 9 and for the same cancer, the 70-79 age group suggests an association with the population.

IV. Discussion

The MCA technique enables the analysis and detection of new relations between categories not observed in the literature. It helped to detect that prostate and colorectal cancer have a high incidence in rural areas. Another important finding was the correlation between lung and urinary bladder cancer and urban areas.

During the period presented (2012-2014), the PCR of Lleida registered approximately 6,000 cases of all possible types of cancer. In this study, only the types most frequently diagnosed cancers (see incidence tables in the Github repository [17]) in the region were selected (colorectal, lung, urinary bladder, prostate, breast cancer). These covered a total of 3408 cases. New relationships were found through applying MCA to detect relations with the features used in a health region with a good balance between urban and rural populations. We based this on a preliminary study [28] which concludes that the incidence of some cancers depends more on the area. However, it does not search for relationships geographic areas and cancers. Another starting point is the work presented in [8]. This studied the most important cancers and their relationship with rural and urban areas. Their most important findings were that the incidence of prostate cancer was most significant in rural areas and that of breast cancer, in urban settings. These articles studied the urban-rural incidence but did not use the MCA technique to explore associations between categories of qualitative variables as we do. To understand the application of MCA, we based ourselves on a study [29] about healthy ageing, and one [30] which concluded that bad driving and crashes could be affected

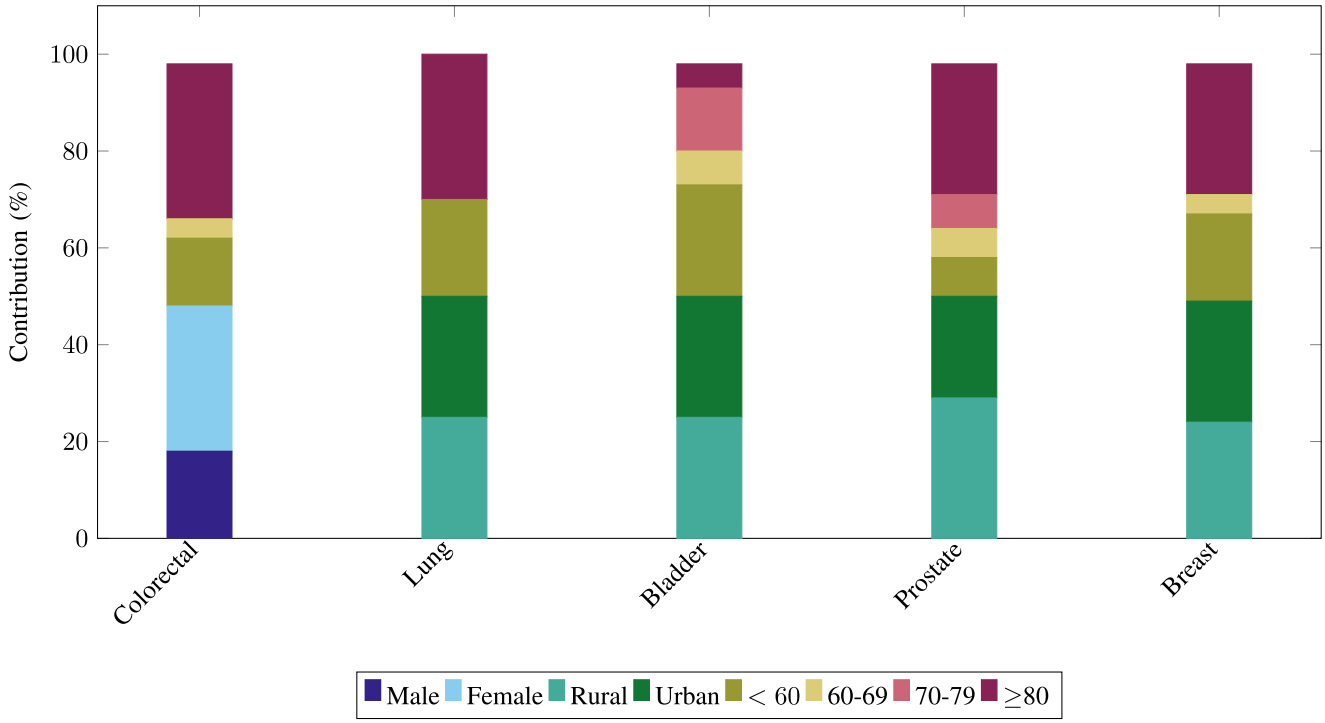

Fig. 8. Contributions by cancer and categories in dimension 1.

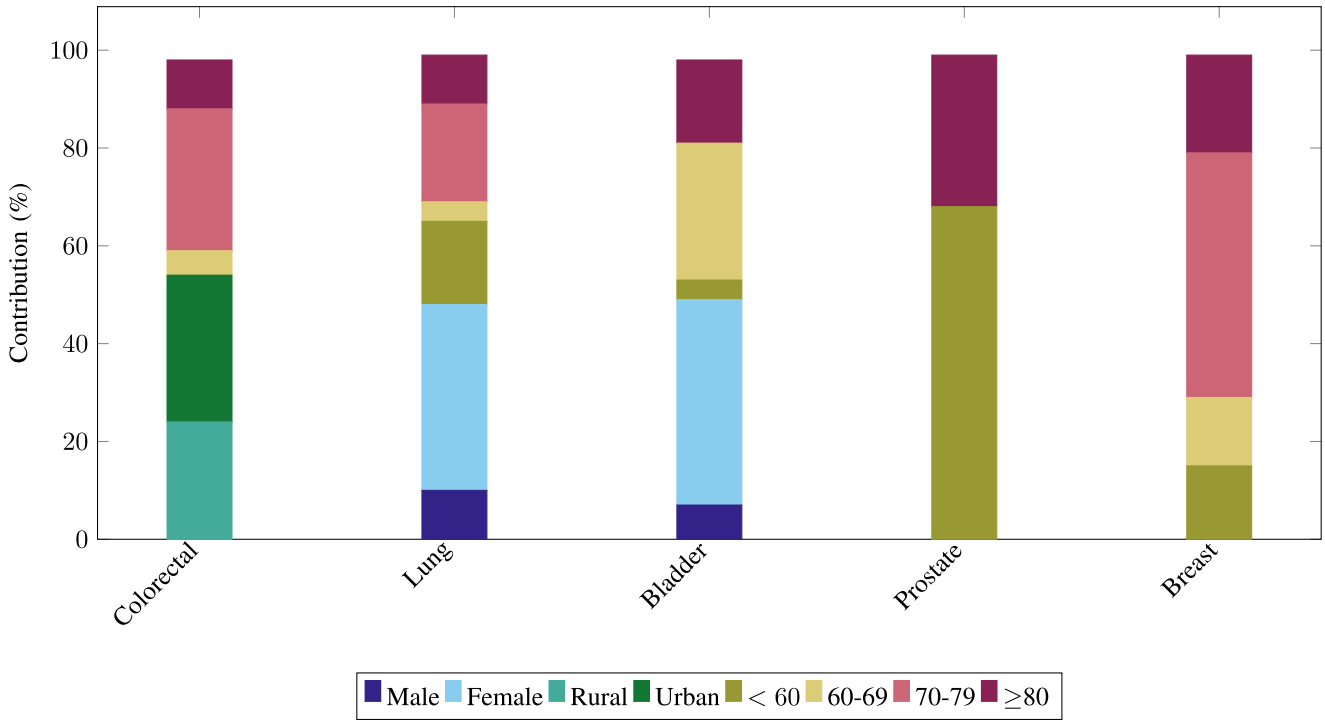

Fig. 9. Contributions by cancer and categories in dimension 2.

by differences between urban and rural areas, traffic volume, driver age and more. In addition, a previous study used MCA to analyse the prognosis in surgery for low rectal cancer [10]. However, to the best of our knowledge no prior studies have used MCA to link types of cancers to rural or urban areas.

Firstly, all the cancers were analyzed together with MCA. The total inertia was 32.9% and the population variance obtained was close to 0 in both dimensions (0.008 in dimension 1 and 0.107 in dimension 2), meaning that this variable combination performed poorly in associating the dataset centered on the population. Next, discarding the gender variable, the total inertia worsened (30.4%), as did the explained population variances (0.062 and 0.035). On also removing the age variable, the population variances improved significantly, to 0.004 and 0.506. This was the best result obtained with the population variable. However, these good results were at the expense of discarding such an important feature as age group.

Some important outcomes were found. These include the lower incidence of colorectal cancer (for either gender) and the moderate rate of prostate cancers among men in rural areas. Males were also more significantly affected by lung and urinary bladder cancer in urban areas. As expected, breast cancer had a high incidence among females. This suggest new hypotheses to deepen and study these specific cancers.

The analysis then studied each cancer separately. Among the population aged <60, **colorectal** cancer affects males severely. On reaching the age of 70-79, this shifted to the urban population. Significant outcomes were obtained in rural females aged ≥ 80. This can be related to the greater age of females in the rural population [31]. These associations with rural populations suggest a high incidence in rural areas. A similar rural incidence was obtained in the study into metropolitan and non-metropolitan areas in the United States [28].

In **lung** cancer, the goodness of the results obtained can be contrasted with human behaviour and genetics. First, the migration of young people influences this in both urban and rural areas. This is seen in females aged <60 in urban areas, contrasting with rural zones, where it tends to affect those aged ≥ 80. In both cases, the red in the picture shows that this associating contribution is very high. These results corroborate the findings of the work presented in [32]. Furthermore, this cancer affects males in the 70-79 age group slightly more. The 60-69 contribution is insignificant in any sense.

The results for **urinary bladder** cancer were similar to those for lung cancer. Urban females aged <60 and then urban males (60-69) contributed moderately, but this decreased even more for the population aged 70-79. Rural females aged ≥ 80 are hardly affected.

As expected, **prostate** cancer only affected males, with a high incidence among the rural population aged <60 and ≥ 80. In contrast the results were insignificant among the other groups in urban environments. The major incidence among those aged <60 in the rural environment is very significant and much attention should be paid to it. In this case, the significant association between rural areas and prostate cancer differs from the incidence of this cancer in other regions analyzed [28]. However, a study into the incidence of cancer in Ireland obtained outcomes that concluded that the risk was higher in rural areas [8].

Again, of course, **breast** cancer only affected females. There was a higher incidence among urban women aged <60 (as with colorectal, lung and urinary bladder cancer). Surprisingly, rural females in the 60-69 age group were hardly affected. The group which contributed the most was those aged 70-79 whatever the population. This contribution was very significant and is a clear example of a case to be studied. Again it affected rural females aged ≥ 80 heavily.

This study has some limitations that should be noted. The postal address registered for each case was where the patient lived at the moment of cancer diagnose. However, this may have changed during the study. Despite this, the number of cases with changed addresses would be very low and this factor is not expected to produce bias in the results. Some lifestyle aspects, such as tobacco and alcohol consumption, profession or other risk factors that could explain some of the differences observed, were not taken into account.

V. Conclusion

There are incipient research efforts to search for correlations between cancers and lifestyle, such as the effect of incidence from living in rural or urban environments. Research using MCA has been applied in various fields, but no one has focused on analysing relationships between cancers and urban and rural lifestyles. This was our main research aim.

Some important outcomes were found, such as the contribution of colorectal cancer (whatever the gender) and prostate cancers among men in rural areas. Also, there was a low incidence of lung cancer but high rate of bladder cancer, especially in urban areas, and the incidence of breast cancer has high in both areas. These outcomes suggest new hypotheses to deepen the study of these specific cancers.

The analysis of each cancer provided additional information. Colorectal cancer severely affected males aged <60, and those in urban areas aged 70-79, as well as women aged ≥ 80 in rural areas. Lung cancer had a high impact on urban females <60, a moderate one on males between 70 and 80 and high again among females aged ≥ 80. Similar but lower results were obtained for urinary bladder cancer. This was moderate in urban females <60 and urban males aged 60-69, decreasing for rural residents aged 70-79 and even more for rural females aged ≥ 80. Prostate cancer, as expected, only affected males. There was a high rate among the rural population aged <60, but this was lower in urban dwellers aged 60-69 and 70-79 before becoming significant again among rural men aged ≥ 80. In contrast, cases of breast cancer were only registered in females in the selected period. Whatever the area, those aged 70-79 were affected the most while the incidence among rural females aged ≥ 80, was somewhat less.

The outcomes obtained help to corroborate suspected trends in several of the relationships detected and stimulate new hypotheses about the risk factors and new techniques to analyse the incidence of cancer. They also help the public health system to focus advice on specific areas and cancers. In future work, it is important to delve deeper into each cancer in order to study its risk factors. This means using new variables, such as tumor characteristics (size, cancer stage or degree of aggressiveness),

treatments, socioeconomic ratio, environmental conditions and mortality. Also, new artificial intelligence algorithms can be explored to search for behavior patterns of cancer, unsupervised clusters or to analyze risk factors and prior patient comorbidities.

References

[1] "World Health Organization. Cancer," Accessed: Jun. 1, 2020. [Online]. Availabel: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] S. Siesling *et al.*, "Uses of cancer registries for public health and clinical research in Europe: Results of the European network of cancer registries survey among 161 population-based cancer registries during 2010–2012," *Eur. J. Cancer*, vol. 51, no. 9, pp. 1039–1049, Jun. 2015.

[3] M. C. White *et al.*, "The history and use of cancer registry data by public health cancer control programs in the United States," *Cancer*, vol. 123, no. Suppl 24, pp. 4969–4976, Dec. 2017.

[4] D. de Salut, *El Càncer a Catalunya*. Monografia 2016 (In catalan), Barcelona: Dept de Salut, 2017, ch. 1, pp. 1–109.

[5] W. E. Zahnd *et al.*, "Rural-urban differences in cancer incidence and trends in the united states," *Cancer Epidemiol., Biomarkers Prevention: A Pub. Amer. Assoc. Cancer Res., cosponsored by the Amer. Soc. Prev. Oncol.*, vol. 27, no. 11, pp. 1265–1274, Nov. 2018.

[6] A. C. Monroe, T. C. Ricketts, and L. A. Savitz, "Cancer in rural versus urban populations: A review," *J. Rural Health*, vol. 8, no. 3, pp. 212–220, 1992.

[7] M. E. O'Neil, S. J. Henley, E. A. Rohan, T. D. Ellington, and M. S. Gallaway, "Lung cancer incidence in nonmetropolitan and metropolitan counties - United States, 2007–2016," *MMWR. Morbidity Mortality Weekly Rep.*, vol. 68, no. 44, pp. 993–998, 2019.

[8] L. Sharp *et al.*, "Risk of several cancers is higher in urban areas after adjusting for socioeconomic status. results from a two-country population-based study of 18 common cancers," *J. Urban Health*, vol. 91, no. 3, pp. 510–525, 2014.

[9] D. Florensa *et al.*, "El registre poblacional de càncer a lleida en zones urbanes i rurals. resultats de l'any 2014." *Butlletí Epidemiol. Catalunya*, vol. 40, no. 12, pp. 252–264, 2020.

[10] R. Mancini *et al.*, "Tumor regression grade after neoadjuvant chemoradiation and surgery for low rectal cancer evaluated by multiple correspondence analysis: Ten years as minimum follow-up," *Clin. Colorectal Cancer*, vol. 17, no. 1, pp. e13–e19, 2018.

[11] I. C. d'Oncologia, "Registre Hospitalari De Tumors ICO/CSUB," [Online]. Available: http://pdo.iconcologia.net/rht/registres.htm

[12] "Idescat, Anuari Estadístic De Catalunya. Densitat De Població. Comarques I Aran, àmbits I Províncies," 2014. [Online]. Available: https://www.idescat.cat/pub/?id=aec&n=249&t=2014

[13] J. García González, "La población rural de España, De los desequilibrios a la sostenibilidad social," *Encrucijadas - Revista Crítica de Ciencias Sociales*, vol. 6, no. 14, pp. 146–149, 2013.

[14] P. Godoy, T. Pedrol, I. Mòdol, and A. Salud, "El registre poblacional de càncer a lleida: Resultats I perspectives," *Butlletí Epidemiològic Catalunya*, vol. 37, no. 7, pp. 161–172, 2016.

[15] P. Godoy-Garcia, T. Pedrol, I. Mòdol-Pena, and A. Salud, "El registre poblacional de càncer a lleida: Resultats de l'any 2013," *Butlletí Epidemiològic Catalunya*, vol. 39, no. 1, pp. 1–11, 2018.

[16] S. K. Kwak and J. H. Kim, "Statistical data preparation: Management of missing values and outliers," *Korean J. Anesthesiol.*, vol. 70, no. 4, pp. 407–411, 2017.

[17] D. Florensa, P. Godoy, J. Mateo, and F. Solsona, "Github repository - A new proposal for analysing cancer in urban versus rural populations," [Online]. Available: https://github.com/didacflorensa/MCA-Cancer

[18] F. Murtagh, "Multiple correspondence analysis and related methods," *Psychometrika*, vol. 72, no. 2, pp. 275–277, 2007, doi: 10.1007/s11336-006-1579-x.

[19] F. Husson and J. Josse, *Multiple Correspondence Analysis*, Boca Raton: Chapman and Hall, Jan. 2014, ch. 11, pp. 165–184.

[20] G. D. Franco, "Multiple correspondence analysis: one only or several techniques?," *Qual. Quantity*, vol. 50, no. 3, pp. 1299–1315, 2016.

[21] C. E. Heckler, "Applied multivariate statistical analysis," *Technometrics*, vol. 47, no. 4, pp. 517–518, 2005.

[22] J. Pagès, "Multiple factor analysis: Main features and application to sensory data," *Revista Colombiana de Estadistica*, vol. 27, no. 1, pp. 1–22, 2004.

[23] M. Greenacre, *Correspondence Analysis in Practice*. Boca Raton: Chapman and Hall, ch. 6, pp. 56–62, 2017, doi: 10.1201/9781315369983.

[24] B. L. Roux and H. Rouanet, *Geometric data analysis: From correspondence analysis to structured data analysis*, Springer, 2005, ch. 1, pp. 18-24, doi: 10.1007/1-4020-2236-0.

[25] R Core Team, *R: A. Language and Environment for Statistical Computing, R. Foundation for Statistical Computing*, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[26] S. Lê, J. Josse, and F. Husson, "Factominer: An R package for multivariate analysis," *J. Statist. Softw., Articles*, vol. 25, no. 1, pp. 1–18, 2008.

[27] A. J. Abdelwahab Yousef, "Male breast cancer: Epidemiology and risk factors," *Seminars Oncol.*, vol. 44, no. 4, pp. 267–272, 2017.

[28] S. Henley, R. Anderson, C.C. Thomas, G.M. Massetti and B. Peaker, "Invasive cancer incidence, 2004–2013, and deaths, 2006–2015, in non-metropolitan metropolitan counties - United States," *MMWR Surveill Summ*, vol. 66, no. SS-14, pp. 1–13, 2017.

[29] P. S. Costa, N. C. Santos, P. Cunha, J. Cotter, and N. Sousa, "The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing," *J. Aging Res.*, vol. 72, no. 2, pp. 257–284, 2013.

[30] S. Das, R. Avelar, K. Dixon, and X. Sun, "Investigation on the wrong way driving crash patterns using multiple correspondence analysis," *Accident Anal. Prevention*, vol. 111, no. 2018, pp. 43–55, 2018.

[31] L. C. M. Fernández and J. M. D. Urrecho, "Envejecimiento Y desequilibrios poblacionales en las regiones españolas con desafíos demográficos," *Ería*, vol. 37, no. 1, pp. 21–43, 2017.

[32] C. D. Viñas, "Depopulation processes in european rural areas: A case study of cantabria (Spain)," *Eur. Countryside*, vo. 11, no. 3, pp. 341–369, 2019, doi: 10.2478/euco-2019-0021.