CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17500

Work Realities and Behavioral Risk Factors in Italy

Angela Andreella, Stefano Campostrini
Department of Economics, Ca' Foscari University of Venice, Italy.

How to cite: Andreella, A.; Campostrini, S. 2024. Work Realities and Behavioral Risk Factors in Italy. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17500

Abstract

The connection between health, work environment, and job characteristics is a relevant issue in public health. However, it is often underexplored due to a lack of reliable data. To address this gap, we have delved into the subject using data from an NCDs-risk factor surveillance system (PASSI). We have examined information collected from respondents regarding their occupations relating to risk factors and health status. The proposed analysis employs text mining and cluster approach for categorical variables to identify sub-populations characterized by different socio-economic situations, risk factors, and job types. Although further analyses are needed to explore the potential of this approach better, initial results are promising. They highlight the practical implications of our findings for public health policies. For example, we found that occupations related to the building industry (for males) and healthcare professions (for females) appear to be associated with higher behavioral risk factors, which could inform targeted interventions.

Keywords: *Job, behavioral risk factors, surveillance system PASSI, categorical data, text mining, clustering*

1. Introduction

Behavioral risk factors, such as smoking, inadequate nutrition, excessive alcohol consumption, and physical inactivity, collectively referred to as SNAP, constitute significant contributors to morbidity and mortality. These factors are prevalent in high-income countries and increasingly in lower-income nations (Noble et al., 2015). Their impact extends to developing chronic diseases, the leading causes of global mortality (World Health Organisation, 2014). Recognizing how risk factors coalesce informs effective preventive interventions. This insight targets specific populations for tailored health promotion and emphasizes the potential impact of addressing multiple risk behaviors concurrently, a key to more impactful public health outcomes (Prochaska & Prochaska, 2011).

It is well known in the literature that SNAP risk factors correlate with other aspects of life, including socio-economic status, such as educational level and economic situation (Flaskerud et al., 2012; Minardi et al., 2011). This, in turn, also leads to a relationship between SNAP and the type of job and work environment, acknowledging the profound impact of work characteristics on health. However, the connection between job characteristics and health/SNAP remains underexplored mainly due to the lack of proper data. For that, a predominant focus on job stress as the connecting factor to behavioral risk factors persists in many analyses, often relying on investigations concentrated on specific occupations. Notable examples include the research of Kouvonen et al. 2007 and Nyberg et al. 2013. Kouvonen et al. 2007 examined the relationship between job stress and smoking and alcohol consumption in public sector employees, while Nyberg et al. 2013 studied the association between job strain and cardiovascular disease risk factors. Focusing on Italy, Chiatti et al. 2010, analyzed exclusively the smoking habit, as Ficarra et al. 2011, but focalizing healthcare professionals.

Instead, the aims of this manuscript are: (1) to evaluate the relevance of considering several job-related variables when health topics are examined, (2) to explore how to analyze these job-related variables since an open-ended question is also present, (3) to discern sub-populations characterized by common SNAP and specific occupational types along with socio-economic variables. We want to give an initial overview of additional information related to work realities that policymakers should consider when targeting sub-populations for health prevention. The exploratory analysis proposed here was possible thanks to the availability of data coming from the cross-sectional Italian surveillance system PASSI (Baldissera et al., 2011). This system has been collecting data about lifestyle, behavioral risk factors, socio-demographic information, and self-diagnosed chronic diseases since 2007 with a high response rate, i.e., approximately 85%. Job-related information has always been collected with some further insights only in the last 10 years.

The analysis proposed is divided into several steps. First, text preprocessing is performed on the primary variable since it is an open-ended question where the interviewer asks about the respondent's job. Then, we perform a cluster analysis for mixed data based on medoids and Gower distances (Gower, 1971), considering the behavioral risk factors, socio-demographic variables, job sector, and classification as covariates. Finally, we analyze each cluster separately, understanding which type of job (coming from the first step) is most prevalent in the different clusters.

The paper is organized as follows. Section 2 describes the data analyzed and the related preprocessing steps. Section 3 briefly defines the clustering approach and related dissimilarity matrix. Finally, Section 4 is devoted to the results, while Section 5 is to the discussion.

2. PASSI Data

We analyze data from the Italian surveillance system PASSI, a sample survey that collects information about lifestyles, behavioral risk factors, socio-demographic information, and self-diagnosed non-communicable diseases. The population of reference is Italian adults ages 18 to 69. For additional information, please refer to Baldissera et al.2011 and the following webpage: https://www.epicentro.iss.it/passi/en/english. Our focus centers on the years preceding the COVID-19 pandemic, specifically from 2014 to 2019, during which job-related variables were recollected. Approximately 40% of the observations exhibit missing values concerning these job-related variables, whereas 88% refer to unemployed respondents. We then have a total of 129,100 observations (56% males). The variables related to socioeconomic variables (first two) and behavioral risk factors (last four) analyzed in the cluster analysis are defined in Table 1.

Table 1. Variables analyzed coming from the Italian surveillance system PASSI.

Variable	Description
Educational level	Low: if below high school; high: otherwise
Economic status	No: if the respondent easily meets financial needs; yes: otherwise.
Alcohol	No: never alcohol; yes: otherwise
Activity	Intense; moderate; no activity
Diet	No fruit; 1-2 portions; 3-4 portions; 5+ portions (per day)
Smoke	Smoker; ex-smoker; never smoke

Furthermore, we examine three variables related to the respondents' employment, defined below. The first stems from the query: "Can you tell me what you do for a living?". The second involves the classification of the declared job according to ISTAT (the Italian National Institute for Statistics) coding (https://professioni.istat.it/sistemainformativoprofessioni/cp2011/). Here, interviewers compile the information directly without querying respondents. As the response to the first job-related question is open-ended, an initial step involves proper text preprocessing to assess its coherence with the ISTAT job class declared by interviewers. Text preprocessing was conducted using the TextWiller R package (Solari et al., 2019), tailored for the Italian language. So, we perform web scraping on the ISTAT website (https://professioni.istat.it/ sistemainformativoprofessioni/ cp2011/) to correlate each job declared by the respondent with the corresponding ISTAT classification. We observed ≈ 60% coherence between the web scraping results and the class declared by the interviewer. After reviewing half of the mismatches, we consider the web scraping results more reliable and utilize them in the clustering step. Finally, the third job-related variable delineates nineteen occupational sectors: agriculture, industry (with specific subcategories, i.e., food, mechanical engineering, electrical and electronic, textile, chemical and ceramics, wood and paper, other), construction, energy-gas-

water-telecommunications, commerce and public establishments, transportation, banks and insurance, school-university, healthcare, public administration, business services, personal services, and law enforcement.

3. Cluster analysis

The joint analysis of several components related to SNAP risk factors at the same time presents challenges. Techniques like Principal Component Analysis (PCA) and factor analysis are commonly used to address high dimensionality. However, using PCA in a high-dimensional and heterogeneous variable set may hinder result interpretation. Factor analysis has limitations with categorical variables, such as the one analyzed in this manuscript. Instead, we opted for cluster analysis to group subjects with similar lifestyles and socioeconomic patterns.

We chose the k-medoids approach (Hastie et al., 2009) instead of other clustering methods, such as k-means and hierarchical clustering, since it works with any type of dissimilarity matrix and is robust with respect to the presence of outliers and noise. It is also less sensitive to initial cluster centers and the assumption of spherical cluster shapes.

As already pointed out, the variables under examination are qualitative ones. Therefore, the Gower distance (Gower, 1971) is utilized in the clustering approach. Let $j = 1, \dots, 9$ the index specifying the covariates defined in Section 2 plus the ISTAT and sector variables and $i, i' \in \{1, \dots, n\}$ the index denoting the respondents. Then, X_{ij} defines the value of the variable j at observation i. The Gower distance matrix $D \in \mathbb{R}^{n \times n}$ has elements $d_{ii'}$ defined as the average of single dissimilarities $d_{ii'j}$ (one for each variable). The single $d_{ii'j}$ is defined depending on the characteristic of the qualitative variable. We consider the educational level, economic status, alcohol variables as asymmetric binary variable while the physical activity, smoke and diet variables are defined as ordered categorical variables and the two job-related variables as categorical ones. Considering $d_{ii'j} = 1 - s_{ii'j}$, if the variable is asymmetric binary, we have $s_{ii'j}$ equals

$$s_{ii'j} = \begin{Bmatrix} 1 & if\ X_{ij} = X_{i'j} = 1 \\ 0 & otherwise \end{Bmatrix} \tag{1}$$

while if the variable is symmetric binary or nominal (i.e., more than 2 categories) we have

$$s_{ii'j} = \begin{Bmatrix} 1 & if\ X_{ij} = X_{i'j} \\ 0 & otherwise \end{Bmatrix}. \tag{2}$$

Finally, if the variables are ordered categorical variables, we have (Podani, 1999):

$$d_{ii'j} = \frac{|r_{ij} - r_{i'j}|}{\max(r_{ij}) - \min(r_{ij})}. \tag{3}$$

where r_{ij} denotes the rank of X_{ij}. Having the dissimilarity $D \in \mathbb{R}^{n \times n}$ the k-medoids clustering approach find $k \in \{1, \dots, n\}$ group of observations minimizing the sum of pairwise dissimilarities within the clusters.

4. Results

We imposed an a priori minimum number of three clusters to avoid reducing the complexity of the observations into a binary category. The silhouette index (Hastie et al., 2009) is used as an internal validation index to estimate the optimal number of clusters. Two clustering analyses were performed, one for the female population and one for the male population. The respondents were selected from among those between 29 and 50 years old.

4.1. Male population

The silhouette index equals 0.12, i.e., 5 is the optimal number of clusters. Figure 1 shows the bar plots for each variable defined in Table 1, while we comment below some of them in detail.

The cluster with the lowest risk of SNAP is the first one characterized by intellectual, scientific, and highly skilled professions (i.e., ISTAT code 2 and labor sector named "services to people"). Looking at the open-ended question, the five most frequent jobs are engineer, teacher, lawyer, office employee, and doctor. Clusters 2 and 4 are characterized by the same job ISTAT classification (laborers, artisans, and farmers) even if they present different scenarios in terms of SNAP and socio-economic variables. However, looking at the labor sector variable and the open-ended question, we found that the cluster with the worst situation is the construction sector, while cluster 4 is characterized by blue-collar and skilled workers in the engineering sector. This is an example of how the combination of the three job-related variables analyzed can give insight into the connection between work environment and health.

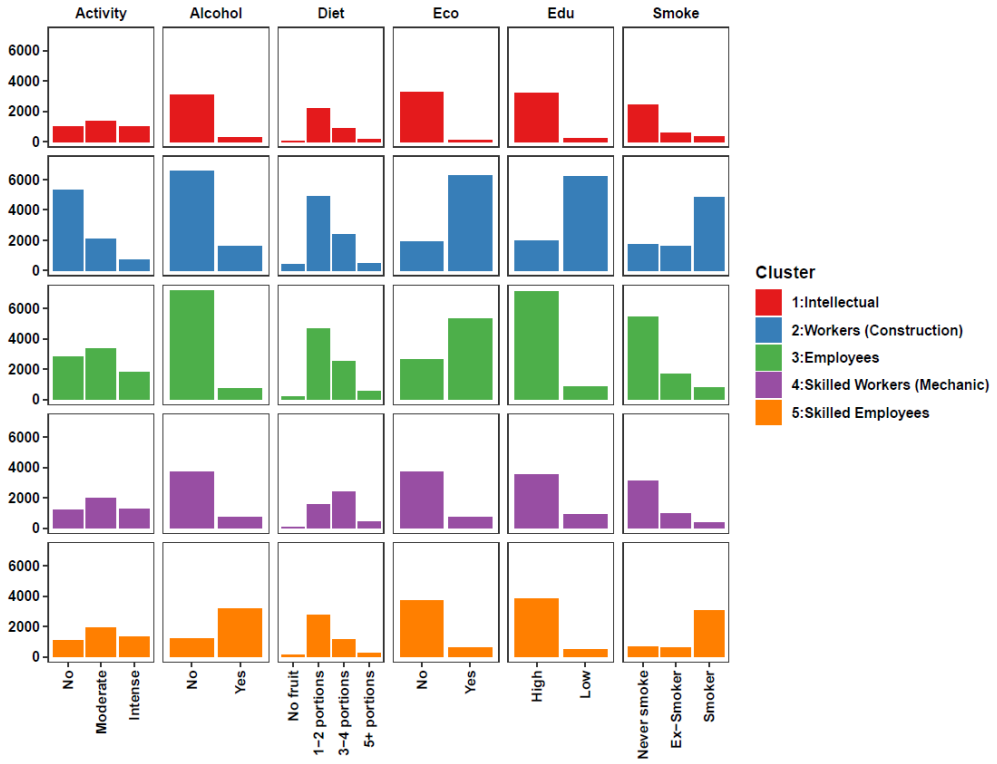

Figure 1. Frequency bar plots of variables defined in Table 1 in each cluster for the male population.

4.2. Female population

The optimal number of clusters is 8 (silhouette index equals 0.14). As in the previous subsection, we report in Figure 2 the bar plots for each variable defined in Table 1, while some clusters are examined in detail, analyzing the three job-related variables and the resulting medoids.

Clusters 1, 2, and 7 emerge as particularly vulnerable. The first cluster predominantly features roles within businesses (e.g., cashiers and shop assistants) and is associated with lower education levels, smoking, and limited physical activity. In contrast, the second cluster is centered around specialized healthcare professions, marked by both smoking and alcohol consumption. The seventh cluster is linked to public administration and predominantly comprises office workers, with absence of physical activity being the primary risk factor. Analyzing the female sub-population proves challenging because 75% of declared jobs are categorized as "employees." However, the other two job-related variables, ISTAT and sector, offer additional insights into the variability within this job type. For instance, clusters 5 and 7 feature the "office employee" job, with the former being associated with no economic problem

and the latter linked to the public administration sector and economic problem. This underscores again the significance of considering all three job-related variables in future studies.

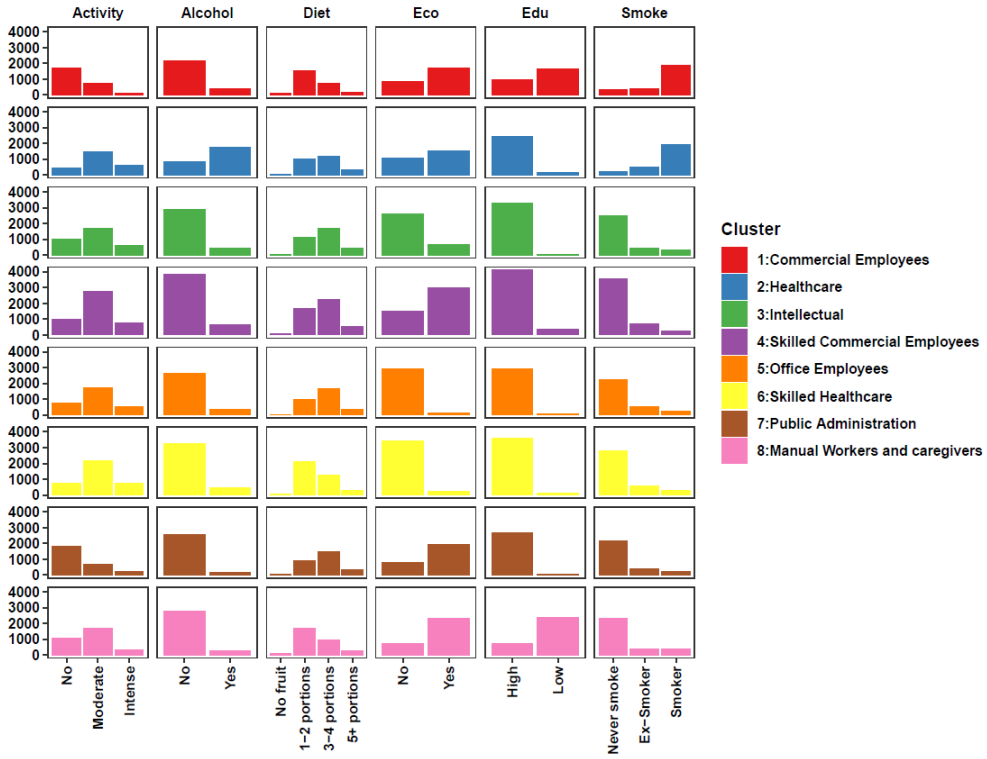

Figure 2. Frequency bar plots of variables defined in Table 1 in each cluster for the female population.

5. Discussion

This study explores the link between behavioral risk factors and socio-economic variables, focusing on detailed job types in Italy. The use of the Italian surveillance system PASSI data seems promising, offering valuable information (when adequately analyzed) that is difficult to gather in other ways. Here, we can analyze three job-related variables (i.e., the specific job coming from an open-ended question, ISTAT classification, and occupational sector), giving different insights into Italy's work realities. We perform a cluster analysis to identify sub-populations with distinct characteristics. Results reveal associations between certain occupations, such as building industry roles for males and healthcare professions for females, and higher levels of behavioral risk factors. The study emphasizes the importance of considering

the three job-related variables in understanding behavioral risk, providing valuable information for targeted health interventions based on specific populations. In particular, thanks to proper text preprocessing and web scraping, the open-ended question gives valuable detailed insights into the job realities of the sub-populations defined by the cluster analysis. However, some criticalities remain and need further investigation. In particular, in the female population, the "employee" job (coming from the open-ended question) is the most prevalent one as well as the most variable in terms of socio-economic work reality. Besides limitations and the need for further analyses, this approach fills the gap of information on health and work environments in Italy, offering this as an example also for other countries in which similar Risk Factors Surveillance Systems are running.

References

Baldissera S, Campostrini S, Binkin N, Minardi V, Minelli G, Ferrante G, Salmaso S. Features and initial assessment of the Italian behavioral risk factor surveillance system (PASSI), 2007–2008. Prev Chron Dis 2011;8(1).

Chiatti, C., Piat, S. C., Federico, B., Capelli, G., Di Stanislao, F., Di Giovanni, P., ... & Manzoli, L. (2010). Cigarette smoking in young-adult workers: a cross-sectional analysis from Abruzzo, Italy. Italian Journal of Public Health, 7(3).

Ficarra, M. G., Gualano, M. R., Capizzi, S., Siliquini, R., Liguori, G., Manzoli, L., ... & La Torre, G. (2011). Tobacco use prevalence, knowledge, and attitudes among Italian hospital healthcare professionals. *European journal of public health*, 21(1), 29-34.

Flaskerud, J. H., DeLilly, C. R., & Flaskerud, J. H. (2012). Social determinants of health status. *Issues in mental health nursing*, 33(7), 494-497.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: Springer.

Kouvonen, A., Kivimäki, M., Väänänen, A., Heponiemi, T., Elovainio, M., Ala-Mursula, L., ... & Vahtera, J. (2007). Job strain and adverse health behaviors: the Finnish Public Sector Study. *Journal of occupational and environmental medicine*, 49(1), 68-74.

Minardi, V., Campostrini, S., Carrozzi, G., Minelli, G., & Salmaso, S. (2011). Social determinants effects from the Italian risk factor surveillance system PASSI. *International journal of public health*, 56, 359-366.

Noble, N., Paul, C., Turon, H., & Oldmeadow, C. (2015). Which modifiable health risk behaviours are related? A systematic review of the clustering of Smoking, Nutrition, Alcohol and Physical activity ('SNAP') health risk factors. *Preventive medicine*, 81, 16-41.

Nyberg ST, Fransson EI, Heikkilä K, Alfredsson L, Casini A, et al. (2013) Job strain and cardiovascular disease risk factors: meta-analysis of individual-participant data from 47,000 men and women. PloS one, 8(6), e67323.

Prochaska, J. J., & Prochaska, J. O. (2011). A review of multiple health behavior change interventions for primary prevention. *American journal of lifestyle medicine*, 5(3), 208-221.

Solari, D., Sciandra, A., & Finos, L. (2019). TextWiller: Collection of functions for text mining, specially devoted to the Italian language. *Journal of Open Source Software*, 4(41), 1256-1257.

World Health Organization. (2014). Global status report on noncommunicable diseases 2014 (No. WHO/NMH/NVI/15.1). World Health Organization.