The Last Social Network: Machine Learning Insights into Online Obituaries and Longevity

Pietro Violo*, Nadine Ouellette*

EXTENDED ABSTRACT

September 2024

Submission for presentation at the
International Population Conference 2025
Brisbane, Australia, July 13-18

*Department of Demography, Université de Montréal

Abstract

This study builds upon our previous work showing that mortality data derived from obituaries accurately represent the general population, at least in Quebec, Canada (Violo and Ouellette, 2024). Using advanced machine learning methods, we aim to extract demographic variables, including marital status and imminent social network size, from a large dataset of 207,572 French-language obituaries. While regular expressions initially captured gender and age at death for 71% of cases, machine learning models significantly improved this to 100% by identifying patterns missed by traditional methods. Our preliminary findings reveal distinct differences in age at death across marital statuses, with single and divorced individuals showing a wider distribution, particularly among men. In contrast, cohabitating women tend to die younger, while married women uniquely show a lifespan disadvantage compared to men, potentially due to health conditions at a younger age. Through DBSCAN clustering, we explore the potential influence of social network size on longevity. This approach promises new insights into the social determinants of mortality, including the impact of family structure, marital status, and social isolation.

Keywords: Longevity | Mortality | Pre-Death Social Ties | Computational Demography | Web Scraping | Data Mining | Machine Learning

Background

We have previously established that mortality schedules derived from online obituaries are representative of the general population, at least for the province of Quebec in Canada (Violo and Ouellette, 2024). From obituary texts, we successfully extracted gender and age at death using pattern matching algorithms known as Regular Expressions. However, we were constrained by the rigidity of Regular Expressions, which requires an exact match between the patterns and the format of demographic information in obituaries for accurate results. In practice, this information is conveyed in countless different ways. The aim of this project is to employ state-of-the-art machine learning methods to extract demographic variables, including those previously mentioned, along with the deceased's marital status and social networks mentioned in bereavement. With these additional variables, we aim to implement a standardized method to fully exploit the richness of online obituaries and gain unprecedented insights into how social networks influence longevity.

Recently, Ebeling et al. (2023) have shown that deaths in developed countries, such as Sweden, most often involve intensive care, a need that increases with age. This result gives us a perspective on how people die, and the circumstances surrounding death. However, for a more comprehensive picture, it is equally essential to understand the social circle of individuals at the time of their death. Obituaries, when properly analyzed, can reveal the social network surrounding individuals at their death, thus providing valuable insights into their end-of-life experiences. De Vries and Rutherford (2004) analyzed a sample of online memorials and found that they are generally written by the children (33%), friends (15%), grandchildren (11%), parents (10%), siblings (8%), and spouses (4%) of the deceased. Although in our case, obituaries are primarily written by funeral homes, this finding underscores the importance of social networks in commemoration, both in the real world and in the digital world. For this reason, it is possible that people with restricted social networks die sooner, as loneliness may play a significant role in health disparities (Raymo and Wang, 2022). Indeed, multiple social mechanisms may influence an individual's longevity. For example, within the immediate family circle, a protective effect of marriage on mortality has been observed, although this is debated (Jaffe et al. 2007; Espinosa and Evans 2008). In the extended family circle, it has recently been suggested that the United States will see an increase in adults without support networks, making these individuals more vulnerable to chronic diseases (Carney et al., 2016). Low social support has been associated with poor physical and psychological health and an increased risk of death among the elderly (Blazer, 1982), and a significant determinant of this loneliness is the absence of children (Carney et al., 2016). This issue becomes even more serious when we consider that the rate of women without children continues to decline,

having recently reached a historic low in Europe (Beaujouan et al., 2017). It has been shown that in 68% of cases, grandparents who are the primary caregivers for their grandchildren experience a deterioration in their health or well-being; conversely, the involvement of non-cohabiting grandparents is associated with health-wise improvements (Danielsbacka et al., 2022). All these social mechanisms have important implications for the longevity of individuals. Taking these points into consideration, we ask: In deaths recorded in online obituaries in Quebec, Canada between 2017 and 2023, do the size of the immediate social network, as well as the marital status, influence longevity?

Data and Methods

The data was obtained from the Necrocanada archive, which compiles obituaries from across Canada on a daily basis. Canadian obituaries are typically presented as text on web pages dedicated to each deceased individual, so we used tools that process the HyperText Markup Language (HTML) of these websites, a technique known as web scraping. Specifically, we used the `rvest` package in `R` (R Core Team, 2022). To filter and retain french obituaries only, we used the `R` package `textcat`. In this way, we collected 207,572 French obituaries from the Canadian province of Quebec between 1 January 2017 and 31 December 2023.

The data extraction process consists of two phases. For the first phase, we extracted gender and marital status using simple machine learning algorithms, such as Random Forest Classification. To accomplish this, we trained a model using the mlr3 package (Lang et al., 2024) in `R`. Our training dataset consists of labeled data previously extracted using regular expressions. The rationale is that training the model on this data would not only cover cases where regular expressions are effective but also capture hidden patterns in the obituary text that would otherwise go undetected.

In the second phase (ongoing), we are fine-tuning more complex machine learning models involving neural networks, such as large language models, to extract both the age at death and the number of relatives left in bereavement. Specifically, we use a Bidirectional Encoder Representations from Transformers (BERT) language model, an open-source machine learning framework for natural language processing (NLP). For our purposes, we plan to use CamemBERT, a language model pre-trained on 138 gigabytes of French text. While this model can be used for various tasks involving French text, we are particularly interested in its named-entity recognition capabilities. This is crucial for identifying key social connections—such as children, grandchildren, great-grandchildren, siblings, and other relatives—mentioned in the obituaries, which helps us map out the immediate social network left behind.

All of our models will be trained on obituary texts from calendar years 2017 to 2022. We will then deploy these models on the 2023 data to evaluate their performance against traditional methods that use regular expressions. This will allow us to assess how accurately demographic variables are extracted through machine learning and whether the models can identify cases that were previously unattainable.

Finally, in addition to using multiple linear regression to identify the determinants of age at death, we will apply Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect emerging clusters. This approach will help us identify patterns and outliers that may reveal different social network structures and how they relate to death. Further details on this can be found in the Next Steps section.

Preliminary Results

From the original dataset of 207,576 French obituaries, we successfully extracted gender, marital status, and age at death for 71% (147,376) of the observations using regular expressions. We then took a random sample of 20,000 from this pool to train a predictive machine learning model based on Random Forest classification for each variable of interest. The model achieved an accuracy of 98.7% for gender, 95.3% for age at death, and 80.5% for marital status on the testing dataset. Building on these results, we were able to increase the retention rate from 71% (via Regular Expressions) to 100% by applying the trained model to the entire dataset (via our predictive machine learning model). We assume that the model's accuracy applies to the observations that were not initially captured by regular expressions. We plan to perform robustness checks to verify the accuracy of this assumption as we develop our third, deep learning model.

Our preliminary results, though still in the early stages, are promising. Both Figure 1 and Table 1 highlight clear differences in the age distribution of deaths in Quebec obituaries based on marital status. Individuals who are single or divorced show a wider age distribution at death, likely because this group is more heterogeneous than others. Within this category, men tend to have a broader age range and die earlier than women. We will investigate further to determine whether this spread is solely due to gender behavioral mortality differences or if being single or divorced is more detrimental to men's health compared to women's. In contrast, women who are cohabitating tend to die younger than their male counterparts, depending on the indicator of interest. The most striking finding concerns married women: they are the only group where women appear to be at a lifespan disadvantage compared to men. This could be attributed to the rarity of such events; typically, death in widowhood is more common among women than men. When a woman dies before her spouse, it is often due to illness at a relatively younger age. It should be noted that we intend to calculate age-specific mortality rates to account for age structure effects, which might alter these preliminary results.

Figure 1: Observed Age-at-Death Distribution by Marital Status and Gender, Quebec, 2018-2023

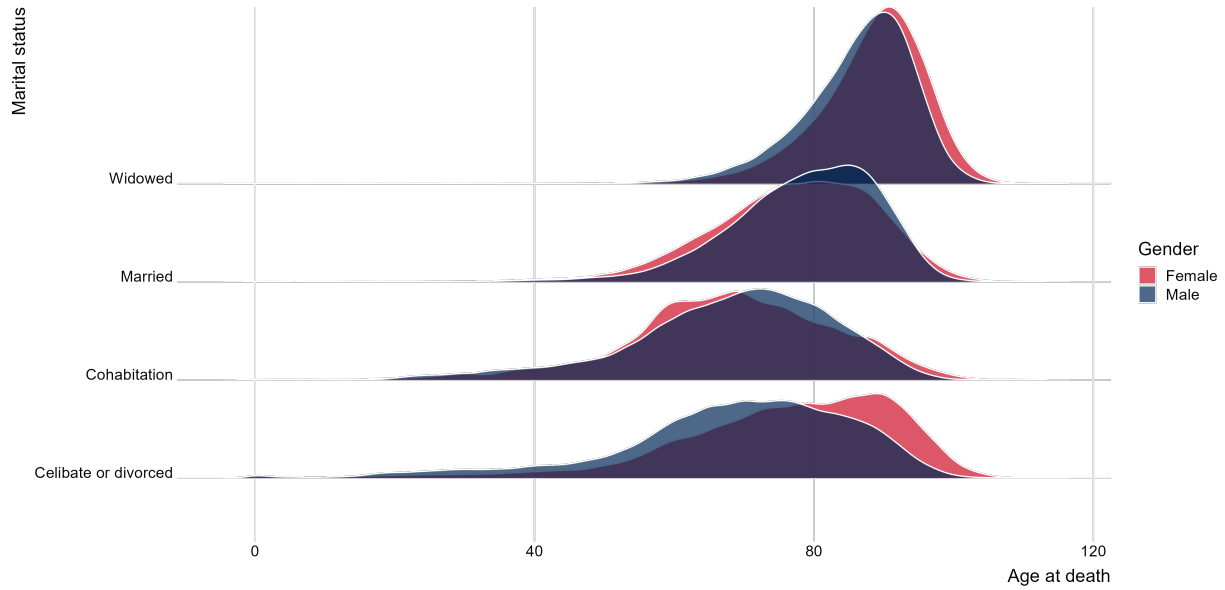


Table 1: Summary measures of Age at Death by Gender and Marital Status, Quebec, 2018-2023

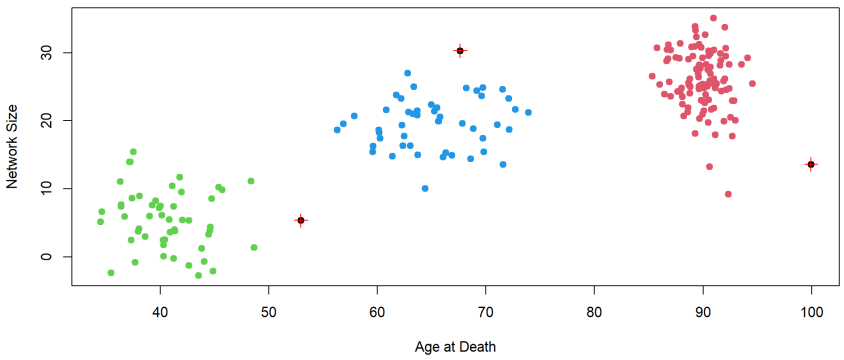
Gender	Marital Status	n	Age at Death			
			Mean	Median	Mode	SD
Woman						
	Single or Divorced	25308	75.87	78.00	90	16.46
	Cohabitation	7861	68.34	69.00	69	14.61
	Married	26893	77.18	78.00	80	11.57
	Widowed	47377	88.25	89.00	90	7.78
Man						
	Single or Divorced	26237	68.75	71.00	70	17.01
	Cohabitation	11692	68.41	70.00	69	14.36
	Married	45196	78.58	80.00	85	10.52
	Widowed	17012	86.31	88.00	91	8.00

Next steps

Our primary objective—understanding the effect of network size on longevity—remains just out of reach, although the completion of our deep learning model is near. To address this question, we will use multiple linear regression models with variables such as the number of children, grandchildren, great-grandchildren, and the presence of a partner, siblings, and parents. Additionally, we will apply DBSCAN clustering, a non-parametric algorithm that groups closely related data points based on variables such as network size, age at death, marital status, obituary length, gender, location, and year of death.

To illustrate this approach, we generated 150 fictional individuals from three different Gaussian distributions based on age at death and network size. The algorithm successfully identified the clusters (Figure 2), which can be interpreted as follows: the green group may represent single and divorced individuals, who tend to die younger and have smaller network sizes; the blue group could represent married individuals; and the red group might correspond to widowed individuals who have reached advanced ages. The stars indicate clear outliers that may warrant individual examination. While this is a fictional example, analyzing the real data could confirm these expected subgroups, reveal contrary patterns, or uncover additional, unforeseen groupings.

Figure 2: DBSCAN Clustering of Age at Death and Social Network Size by Marital Status, Fictional Data



References

Beaujouan, E., T. Sobotka, Z. Brzozowska, and K. Zeman (2017). La proportion de femmes sans enfant a-t-elle atteint un pic en europe ? *Population & Sociétés 540*(1), 1–4.

Blazer, D. G. (1982, May). Social support and mortality in an elderly community population. *American Journal of Epidemiology 115*(5), 684–694.

Carney, M. T., J. Fujiwara, B. E. Emmert, T. A. Liberman, and B. Paris (2016, October 23). Elder orphans hiding in plain sight: A growing vulnerable population. *Current Gerontology and Geriatrics Research*, e4723250.

Danielsbacka, M., L. Křenková, and A. O. Tanskanen (2022, September 1). Grandparenting, health, and well-being: A systematic literature review. *European Journal of Ageing 19*(3), 341–368.

de Vries, B. and J. Rutherford (2004, August 1). Memorializing loved ones on the world wide web. *OMEGA - Journal of Death and Dying 49*(1), 5–26.

Ebeling, M., A. C. Meyer, and K. Modig (2023, July). Variation in end-of-life trajectories in persons aged 70 years and older, sweden, 2018–2020. *American Journal of Public Health 113*(7), 786–794.

Espinosa, J. and W. N. Evans (2008, September 1). Heightened mortality after the death of a spouse: Marriage protection or marriage selection? *Journal of Health Economics 27*(5), 1326–1342.

Jaffe, D. H., O. Manor, Z. Eisenbach, and Y. D. Neumark (2007, July 1). The protective effect of marriage on mortality in a dynamic society. *Annals of Epidemiology 17*(7), 540–547.

Lang, M., Q. Au, S. Coors, P. Schratz, and M. Becker (2024). *mlr3learners: Recommended Learners for 'mlr3'*. R package version 0.7.0, https://github.com/mlr-org/mlr3learners.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Version 4.2.1.

Raymo, J. M. and J. Wang (2022, June 1). Loneliness at older ages in the united states: Lonely life expectancy and the role of loneliness in health disparities. *Demography 59*(3), 921–947.