Who are the loneliest Americans?

Draft

Joe Marlo

2020-11-01

Abstract

Time spent alone has been increasing among Americans. It is hypothesized this increase in alone time has been unequally split amongst the various subpopulations of the United States. Instead of subjectively segmenting the population, clustering of Americans is performed via optimal matching and hierarchical clustering to find natural groupings based on time-use activities. Once the groupings are established, an exploratory analysis of the change in alone time performed.

Contents

Background

Time spent alone has been increasing among Americans. This can have numerous health effects and it may be impacting subpopulations differently. Data from the American Time Use Survey shows the mean amount of time spent on non-work activities with no other person present has steadily increased from ~295min per day to ~330min per day from 2003 to 2018/

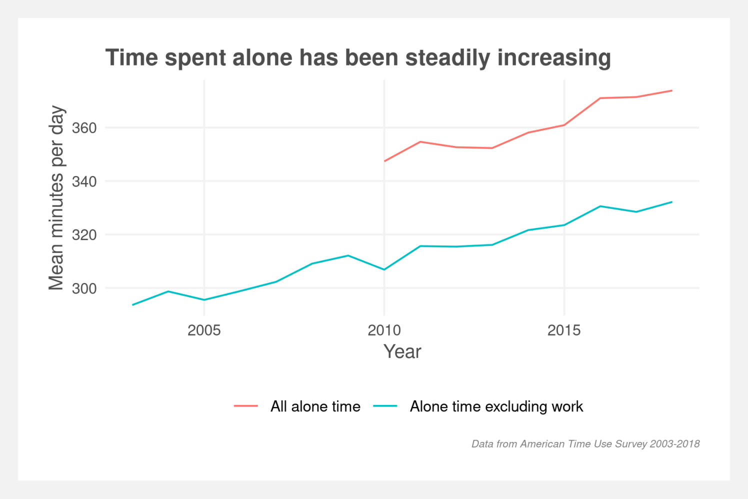

Figure 1: Mean alone time 2003-2018

It's intuitive that this increase may not be evenly distributed across the population. Rather than subdividing the population by demographics, the population can be divided using sequence analysis and unsupervised learning techniques to find clusters of similar time-use patterns. These in turn may represent distinct demographic groups (e.g. a cluster with large amounts of time spent on education consists mostly of sub 25 year olds) but are not direct measurements of demographics. This clustering methodology allows demarcation of groups based on their activity and may capture groups such as students, workers, and the elderly.

Research question

> Are increases in time spent alone equally affecting different subpopulations of Americans?

Loneliness

Traditional literature covering loneliness dates back to at least the 1960s where multiple researchers (Eddy (1961), Sisenwein (1964), Bradley (1969)) applied Likert-like scales to measuring loneliness on unidimensinoal measures. Daniel Russell, et. al (1978) introduced the UCLA Loneliness Scale[1] to formally measure loneliness using a multiple dimensional approach. And finally, a simple approach (Bradburn 1969) measured loneliness simply by asking participants if they feel lonely.

Popular articles in the NY Times (Fountain, 2006[2]), The Atlantic (Marche, 2012[3]), and TIME (Durcharme, 2020[4]) tend to focus on an increase in loneliness among Americans and cite popular statistics such as "most

[1] Russell, D., Peplau, L.A A., & Ferguson, M. L. Developing a measure of loneliness. Journal of Personality Assessment, 1978, 42, 290-294

[2] https://www.nytimes.com/2006/07/02/weekinreview/02fountain.html

[3] https://www.theatlantic.com/magazine/archive/2012/05/is-facebook-making-us-lonely/308930/

[4] https://time.com/5833681/loneliness-covid-19/

adults only have two people they can talk to about the most important subjects in their lives." However, other researchers have found seemingly disparate results. Clark, et. al (2014) performed a meta-analysis of American college students and American high school students using the UCLA Loneliness Scale and found that there have been declines in loneliness from 1978-2009 (Study 1) and 1991-2013 (Study 2).

The results may not be so disparate and could just be confounded by multiple attributes. Tijuis, et. al (1999)[5] sought to investigate if loneliness increases in old age or if it was due to cohort effects. They surveyed 939 men in 1985, 1990, and 1995 regarding a loneliness scale. They found that for their oldest cohort (born 1900-1920), the loneliness scores increased but the scores did not increase for the younger groups. They attributed their increase due to ageing and not due to cohort.

Perhaps most importantly, the risks associated with loneliness. Golden, et. al (2009)[6] also found loneliness increased with age and that wellbeing, depressed mood, and hopelessness were all independently associated with loneliness. Tomaka, et. al (2006)[7] found that "belongingness support related most consistently to health outcomes." Importantly they note that "isolation and loneliness may be related, there is no neccessary relationship between the two".

Alone time definition

The American Time Use Survey tracks alone time via a computation of other collected variables. For each activity — except those noted below — the BLS tracks the number of participants present during the activity. Alone time is tallied only during activities for which only the primary respondent is physically present. Compared to many studies pertaining to loneliness, this approach is measuring the quantity of time spent alone and is not addressing the mental state of loneliness.

The benefit of this approach is detailed data on the length of time and description of the activity in which the person is alone. The shortcomings of this approach is that it only pertains to physicality. Therefore mental health is not directly addressed and activities such as phone or video calls will be labeled as 'alone' unless there is an additionally person physically present. Additionally, a few activities are specifically excluded from the tally including:

- Working
- Sleeping
- Washing, dressing, or grooming
- Personal/private activities
- Any time in which the respondent refused to provide activity detail

The question posed to respondents to define who was present:

> "Who was in the room with you / Who accompanied you?"

The BLS also includes another variable, `TRTALONE_WK` which is similar to `TRTALONE` but includes alone time during work activities. This is excluded for the analysis as it is only available 2010-2018.

[5] Tijhuis, M. A., et al. "Changes in and factors related to loneliness in older men. The Zutphen Elderly Study." Age and ageing 28.5 (1999): 491-495

[6] Golden, Jeannette, et al. "Loneliness, social support networks, mood and wellbeing in community-dwelling elderly." International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences 24.7 (2009): 694-700.

[7] Tomaka, Joe, Sharon Thompson, and Rebecca Palacios. "The relation of social isolation, loneliness, and social support to disease outcomes among the elderly." Journal of aging and health 18.3 (2006): 359-384.

American Time Use Survey

The American Time Use Survey data contains a nationally representative estimates of how, where, and with whom Americans spend their time. The sampling frame is the Census Current Population Survey (CPS) and covers over 200,000 interviews conducted from 2003 to 2018. The survey is phone-based, repeated cross-section, and randomized by day of the week. The data can be linked to the CPS for detailed household demographic information.

Methodology

The analysis consists of two major parts: 1. Calculate and cluster sequences of activities and 2. Model alone time as a function of year and cluster to understand if alone time is increasing or decreasing for a given cluster.

Clustering methods will be used to determine similar sequences of how individuals spend their day irrespective of alone time. The primary techniques will be using optimal matching string editing techniques. This is non-parametric technique that determines distance between two sequences by the least number of operations that are required to convert one sequence into the other. The operations are insertion, deletions, and substitutions where insertion/deletion are typically referred to collectively as 'indel.'

Data detail

Data comes from The American Time Use which surveys how Americans spend their time. The diary (`atusact_0318` file) and CPS (`atuscps_0318` file) data are used from the 2003-2018 Multi-Year Interview dataset. The data details each minute of the respondent's day by mapping it to a list of 465 activities. The author then aggregated these 465 activities into 15 activities based on the Bureau of Labor Statistics' (BLS) hierarchical definitions and the author's judgment. See Appendix Table 3 for the aggregation mapping. Additionally, to reduce computation load, each respondents' day was summarized into 48 thirty-minute windows representing the modal activity during the window, and a weighted sample of 10,000 individuals was chosen.

Time use varies greatly between week and weekend days so only weekdays are included in the analysis. Similarly, holidays are excluded.

The activities of the sampled population are visualized below. Overall, the proportion of individuals participating in a given activity at a given time is mixed. Work (light pink) contains the plurality of individuals during midday, leisure (violet) in late afternoon, and sleep (blue) in the night.

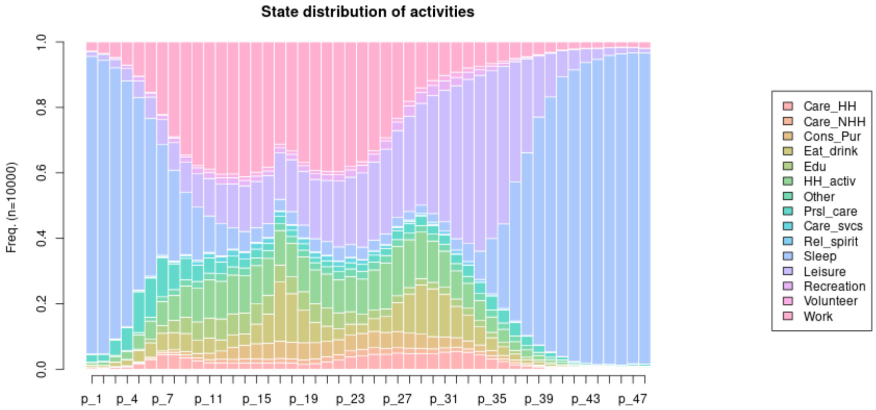

Figure 2: State distribution (p1 represents 4am and every period thereafter is 30 minutes)

Distance measures

String editing techniques will start by aggregating the different types of activities from 465 specific activities into 15 activities based on their hierarchical definitions provided by the BLS. These 15 activities are then be recoded as single character strings representing how an individual spends each 30 minute period of their day. Their full day's activity is represented by the resulting 48 character string.

The distance between the respondents string sequences can be calculated using a number of different string distance measures under the optimal matching framework. First, measures with constant insertion/deletion (indel) and substitution costs will be considered. Then the substitution costs will be calculated using the transition rate between states (TRATE).

Measures to be explored:
- Hamming: only allows substitution
- Longest common subsequence (LCS): only allows insertions and deletions
- Dynamic Hamming: only allows substitutions but costs depend on position within the sequence
- TRATE: costs derived from transition rates

Each of these have advantages and disadvantages that will be explored along with their impact on the final clustering.

The Hamming distance, in this context, implies we are firmly interested in not just the sequence of the time-use activities but also when those activities occur with the day. This aligns with an intuitive understanding that one person waking up in the morning and having breakfast is not an identical activity to another person waking up in the evening from a nap and having dinner even though both are participating in the activity of "sleep" and then "eating."

LCS is the inverse of Hamming. Substitutions are not allowed, and the pattern of interest is sequence of the time-use activities agnostic of when they occur. Referencing on the previous example, the morning and evening persons' activities would be considered identical.

These both may be too stringent of a definition, though. Allowing indels by setting low costs for substitutions may offer a more flexible approach while still retaining the general structure of the day.

Dynamic Hamming is a modification of Hamming which still only allows substitutions but the costs depend on the position within the sequence. The result is slightly more flexibility but is still heavily sensitive to the timing of activities similar to Hamming.

TRATE is fundamentally different. The costs are derived from the observed transition rates between activities. The substitution cost is then calculated via $cval - P(i|j) - P(j|i)$ where $cval$ is a constant and $P(i|j)$ is the probability of transition from state j to i.

Clustering

Once the string distance measures are applied, the resulting distance matrix is clustered using the Ward D2 hierarchical clustering algorithm. The number of clusters is determined using the Silhouette width method which is a measure of distance between clusters. The larger the Silhouette width, the more well separated the clusters.

Efficacy of cross-sectional clustering

The data consists of cross-sectional observations of individuals' time use. The clusters are computed across years. Therefore, no single respondent represents more than one year but individual clusters span multiple years. This may present issues as the metric of interest is the change in time spent alone for a given cluster.

The respondents were split into two groups - one for years 2003-2010 and one for 2011-2018 - and then clustered separately. The resulting groups showed similarity in clusters between these two groups and indicate there will be no discontinuity or heterogeneity issues when clustering across all years.

Final clusters

The final clusters were created using a weighted sample of 10,000 respondents from the 2003-2018 surveys. The number of clusters is determined using the Silhouette method, and each edit distance method agrees that between three and six clusters are the ideal amount.

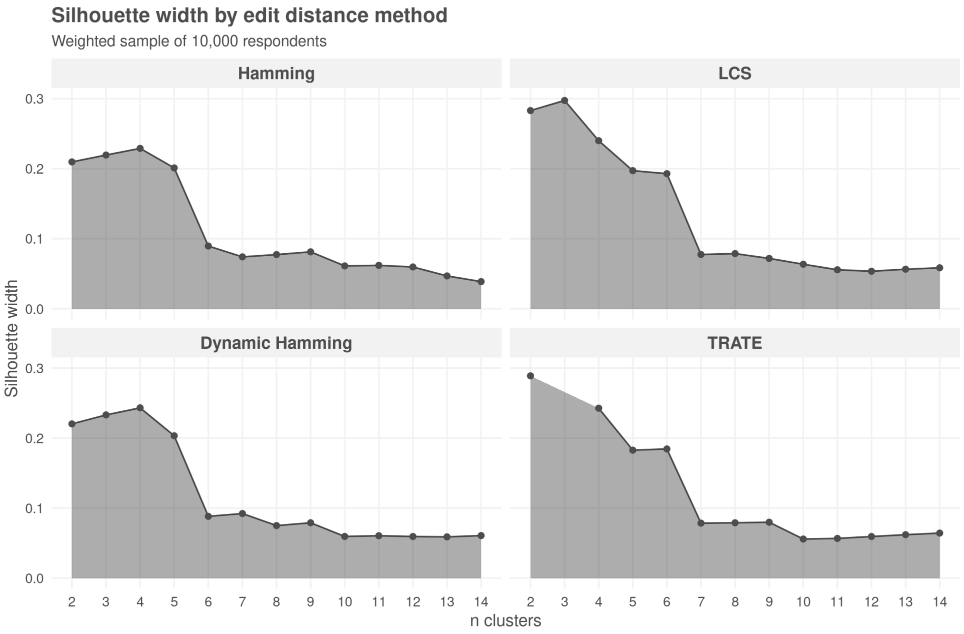

Figure 3: Silhouette comparison

The following dendrograms show the separation when k = 4. Interesting, the dendrograms are similar across the various edit distance methods. As expected, Hamming and Dynamic Hamming are the most similar. LCS and TRATE also appear similar. All four cut the four clusters at roughly the same points indicating there may be a resilient underlying structure to the time-use data.

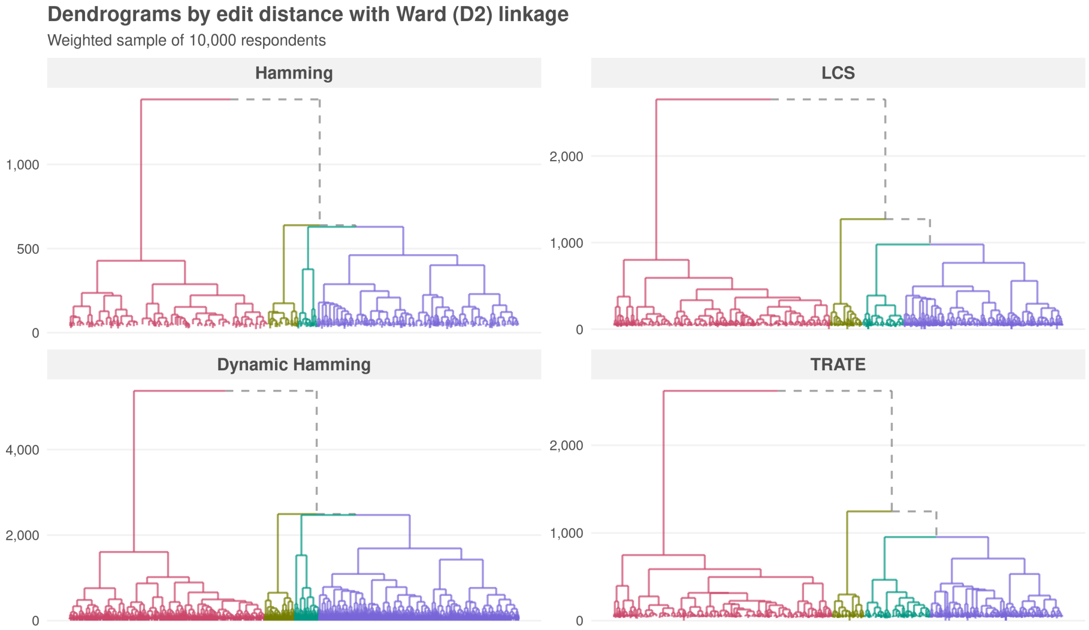

Figure 4: Dendrogram comparison

The following state sequence, distribution, and modal activity plots tell the same story. Each edit distance methods finds a cluster most defined by work (pink), by education (light green), and leisure (violet). Both Hamming variants further find a cluster defined by evening work. Whereas the LCS and TRATE methods appears to split the leisure cluster into housework.

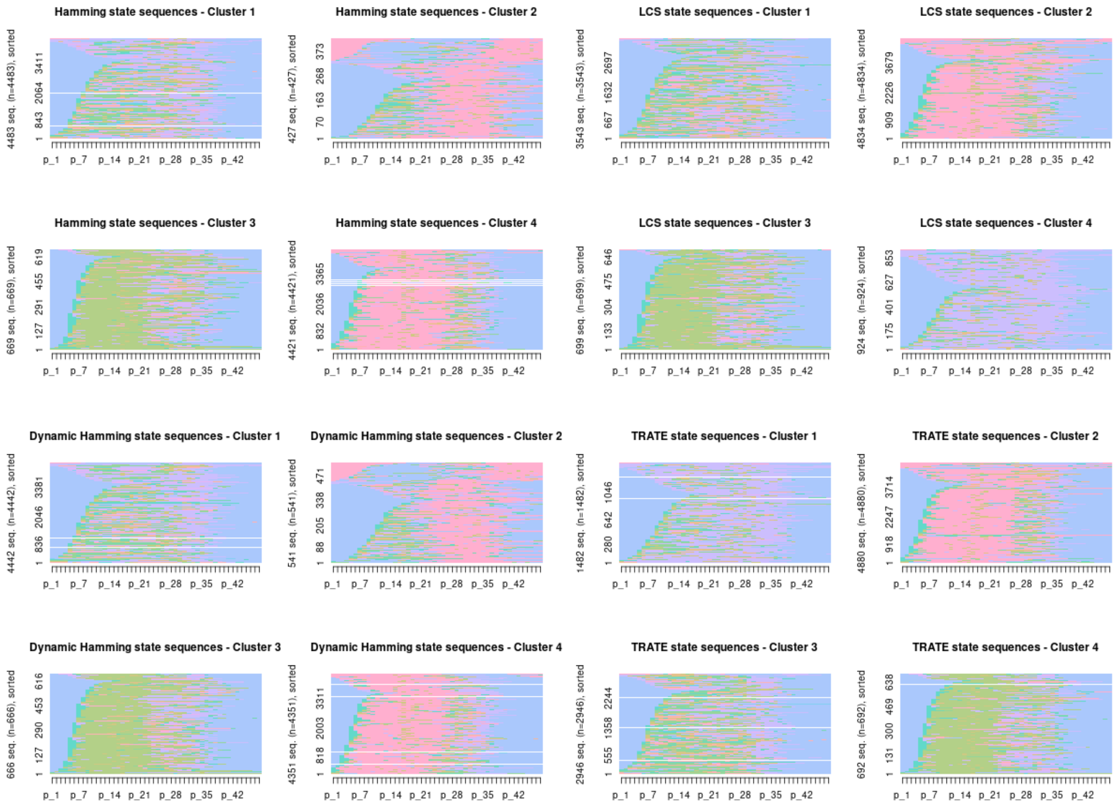

Figure 5: Sequence plots

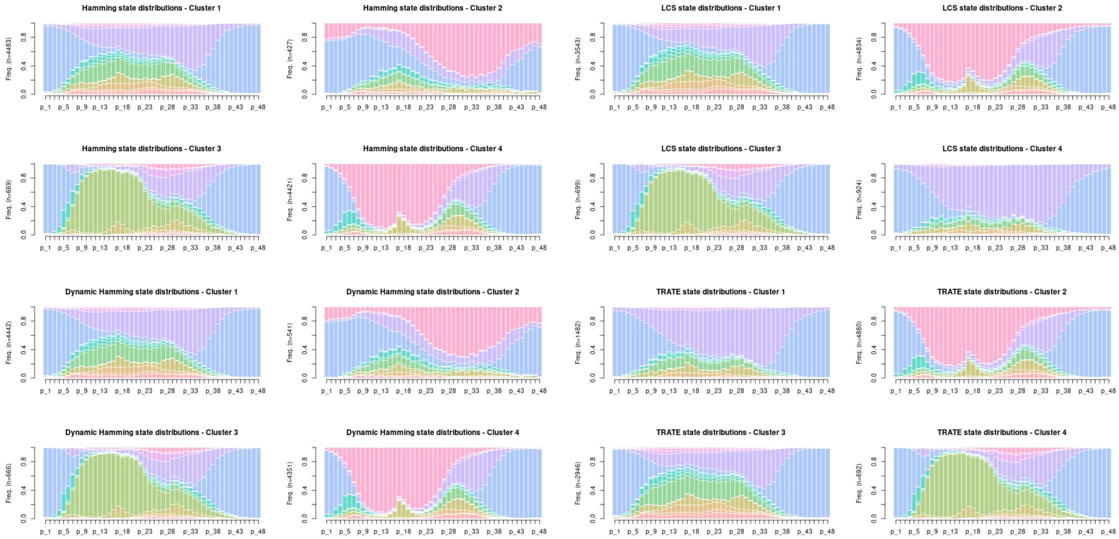

Figure 6: Distribution plots

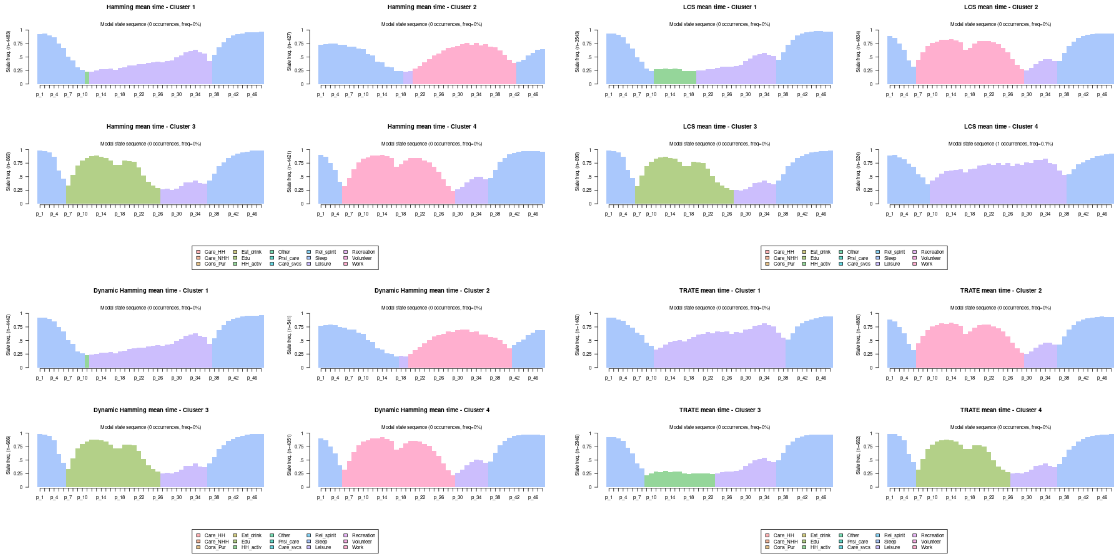

Figure 7: Modal activity plots

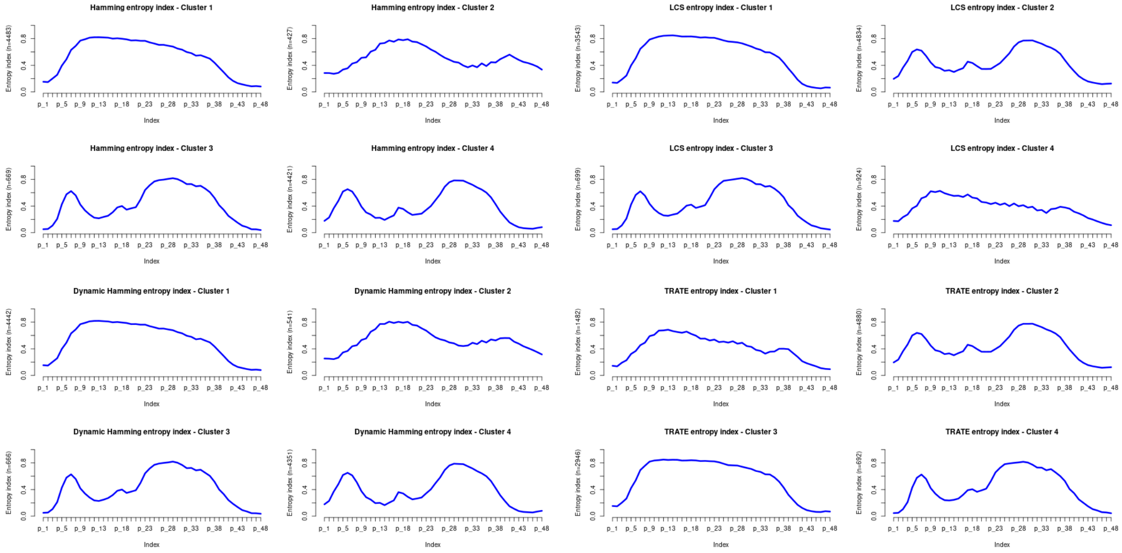

Figure 8: Entropy plots

Categorization of clusters

The sequence and proportion plots illustrate the contents of each cluster are mostly consistent across the pairs of edit distance methods. For brevity, only the Dynamic Hamming clusters will be used in the modeling portion of the analysis. Each cluster can therefore be identified by their common characteristics: one cluster consisting of mostly daytime work will now be labeled as “Day workers”, a second cluster of mostly evening and night work “Night workers”, a third dominated by education “Students”, and then the fourth contains a mixture of activities with no dominate characteristics so this will be referred to as “Uncategorized.”

Modeling

The goal is to understand how alone time has varied across different groups (i.e. clusters). Therefore, the unit of interest is the slope of alone time.

Single-level and multilevel models are both appropriate. Multilevel models "account for individual- and group-level variation in estimating group-level regression coefficients" and "estimate regression coefficients for particular groups" (Gelman Hill 2006). However, the clustering methods are signaling only four clusters are optimal. Gelman and Hill also note that "when the number of groups is small (less than five, say), there is typically not enough information to accurately estimate group-level variation. As a result, multilevel models in this setting typically gain little beyond classical varying-coefficient models."

As such, the approach is to fit both single- and multilevel models and draw comparisons.

Data generating process

The time-use data is inherently count data. Uni-variate densities (below) show that the data is not distinctly Poisson, not zero-inflated, and overdispersion is an issue.

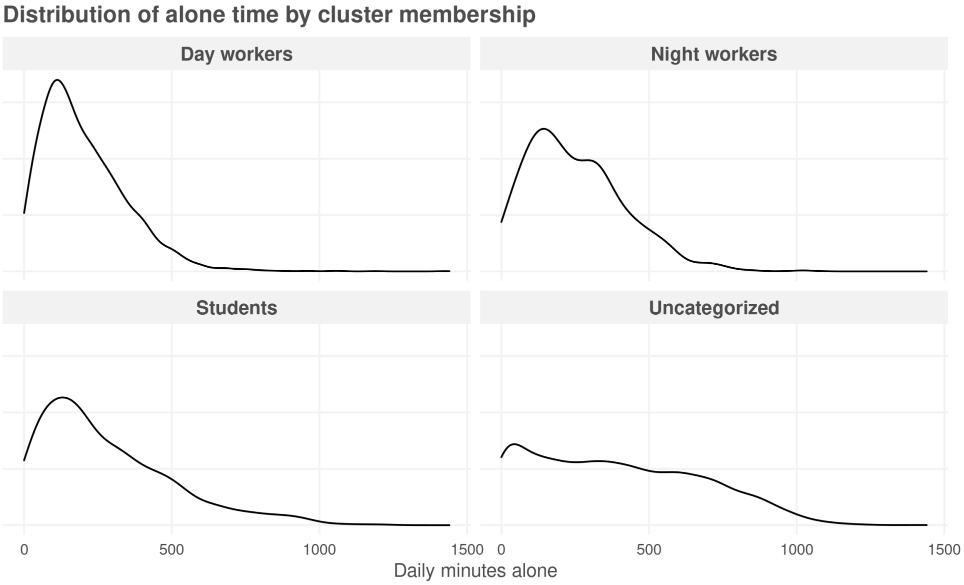

Figure 9: Densities of alone time by cluster

The mean daily alone minutes range from 202 to 400 per each cluster and the variances range from 19,724 to 81,265. This violates an assumption of Poisson models: mean equals variance. It suggests that the standard errors in the Poisson model will be severely underestimated.

Dispersion can also be defined as estimated dispersion via Gelman and Hill (2006, pg 114):

$$
\begin{aligned}
dispersion\, ratio &= \frac{1}{n-k}\sum_{i=1}^{n} z_i^2 \\
n &= \#\, of\ data\, points \\
k &= \#\, of\ predictors \\
z &= standardized\ residuals = \frac{y_i - y^i}{sd(y^i)}
\end{aligned}
$$

Fitting a Poisson multilevel model to this data results in a dispersion ratio of 151.2 — far larger than 1 — indicating a quasi-Poisson or negative binomial model is necessary. Otherwise, the standard errors will need to be corrected by multiplying by a factor of $\sqrt{151} = 12.3$.

Ultimately, the data was transformed using square-root, and the resulting uni-variate densities are "normal enough" that a standard MLM can be fitted.

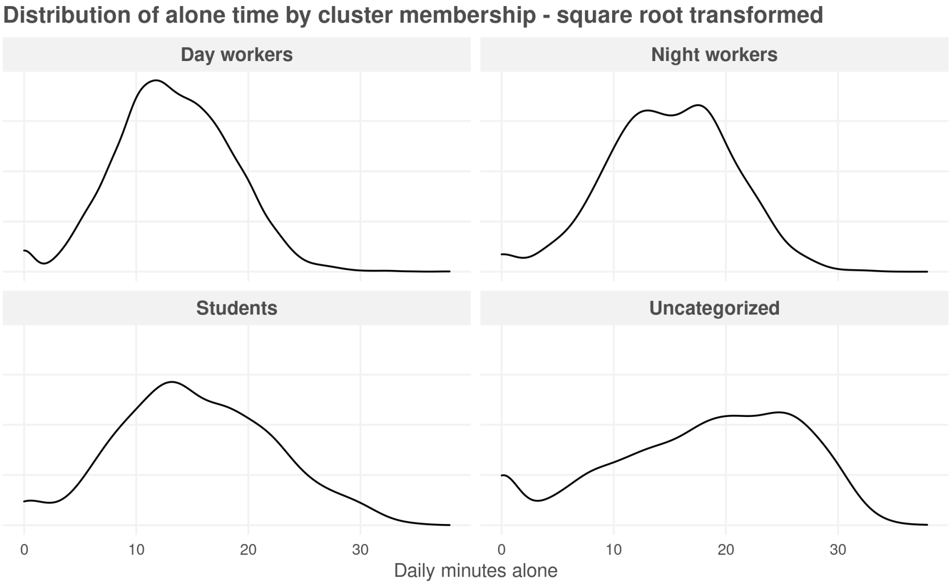

Figure 10: Transformed densities of alone time by cluster

Multiple single-level linear models

Four single-level linear models are fit to the data individually to estimate the effect of `Year` on per cluster per distance method. The only cluster to significantly differ from zero is the "Students" cluster. These point estimates range from -0.25 to -0.01, meaning the mean amount of alone time for students, on average, decreased approximately 0.12 minutes per year from 2003 to 2018.

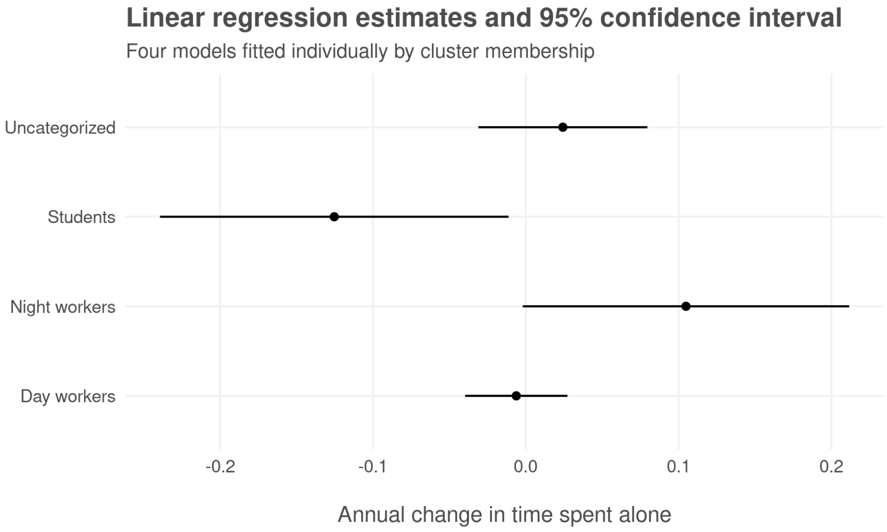

Figure 11: Linear regression estimates

Multilevel models

A multi-level model with cluster as varying intercept and year as fixed and random slope is proposed. However, the model fails to converge because !!!!!

The model form in R syntax:

$$\texttt{Alone time} \sim \texttt{year + (year | cluster)}$$

Associated demographics

The associated demographics of the clusters skew towards expected values. The day workers consist mostly of mid-20 year-olds to 60 year-olds. The night workers skew much younger. And students consists almost entirely of sub 20 year-olds while the Uncategorized cluster is a mixed bag.

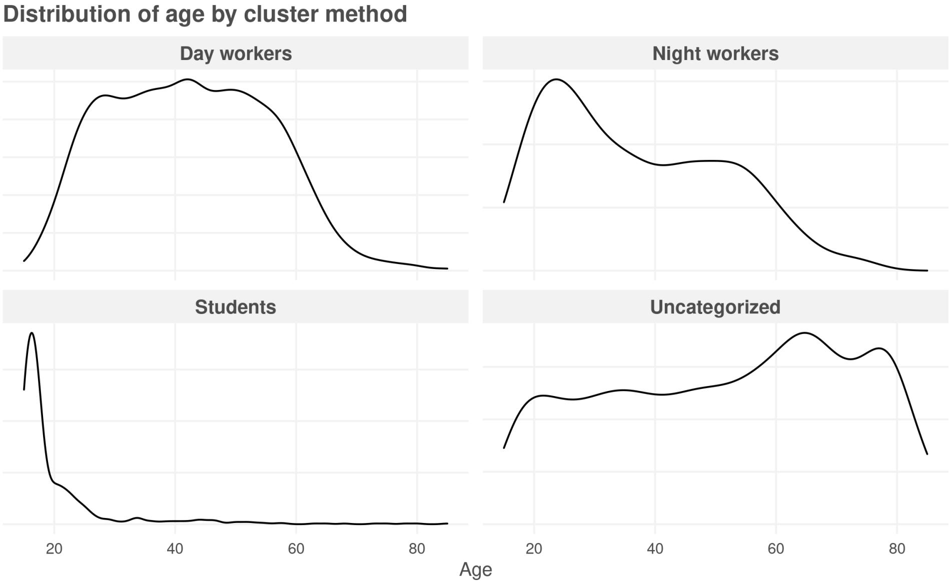

Figure 12: Densities of age by cluster

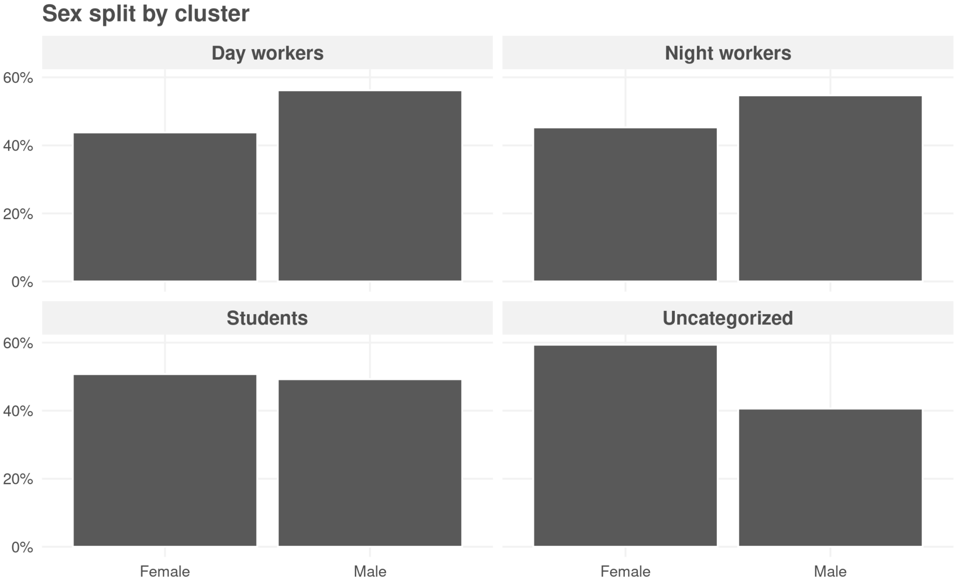

Figure 13: Sex split by cluster

Conclusion

The optimal matching sequence techniques were able to find distinct clusters among the respondents' time-use activities. The four edit-distance methods found similar clusters which may signal a strong underlying structure to American's time-use. Using these resulting clusters as input into modeling proved less than fruitful. Simple OLS linear regression found that the cluster representing students is associated with a statistically significant decline in alone time between 2003 and 2018. However, the effect is minimal at approximately seven seconds per year. Attempts to apply multi-level models was difficult due to convergence issues.

Software

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software, 40(4), 1-37. DOI http://dx.doi.org/10.18637/jss.v040.i04.

Studer, M. & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures, Journal of the Royal Statistical Society, Series A, 179(2), 481-511. DOI http://dx.doi.org/10.1111/rssa.12125

Daniel Müllner (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. Journal of Statistical Software, 53(9), 1-18. URL http://www.jstatsoft.org/v53/i09/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36. URL http://www.jstatsoft.org/v61/i06/.

van der Loo M (2014). "The stringdist package for approximate string matching." *The R Journal*, *6*, 111-122. URL https://CRAN.R-project.org/package=stringdist.

Appendix

Other

Table 1: Activity aggregation mapping

Activity code	Description
t0101.*	Sleep
t010[2-9].*	Personal Care
t019.*	Personal Care
t1801.*	Personal Care
t02.*	Household Activities
t1802.*	Household Activities
t03.*	Caring For Household Member
t1803.*	Caring For Household Member
t04.*	Caring For Nonhousehold Members
t1804.*	Caring For Nonhousehold Members
t05.*	Work
t1805.*	Work
t06.*	Education
t1806.*	Education
t07.*	Consumer Purchases
t1807.*	Consumer Purchases

Activity code	Description
t08.*	Professional & Personal Care Services
t1808.*	Professional & Personal Care Services
t09.*	Other
t1809.*	Other
t10.*	Other
t1810.*	Other
t11.*	Eating and Drinking
t1811.*	Eating and Drinking
t12.*	Socializing, Relaxing, and Leisure
t1812.*	Socializing, Relaxing, and Leisure
t13.*	Sports, Exercise, and Recreation
t1813.*	Sports, Exercise, and Recreation
t14.*	Religious and Spiritual
t1814.*	Religious and Spiritual
t15.*	Volunteer
t1815.*	Volunteer
t16.*	Other
t1816.*	Other
t1818.*	Other
t1819.*	Other
t189.*	Other
t50.*	Other

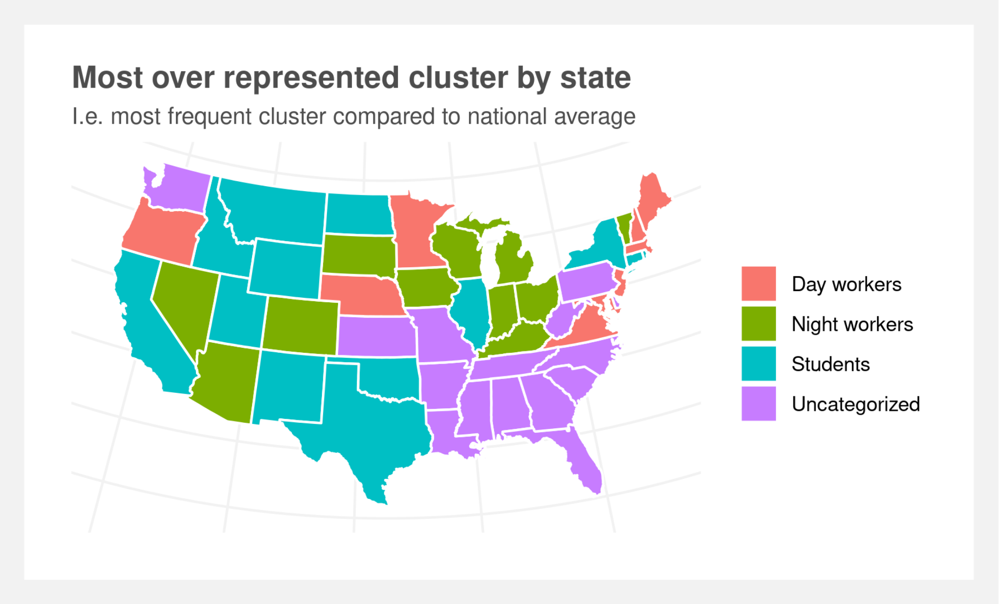
Most over represented cluster by state
I.e. most frequent cluster compared to national average
Day workers
Night workers
Students
Uncategorized

Figure 14: Most over represented cluster by state

Results from previous models

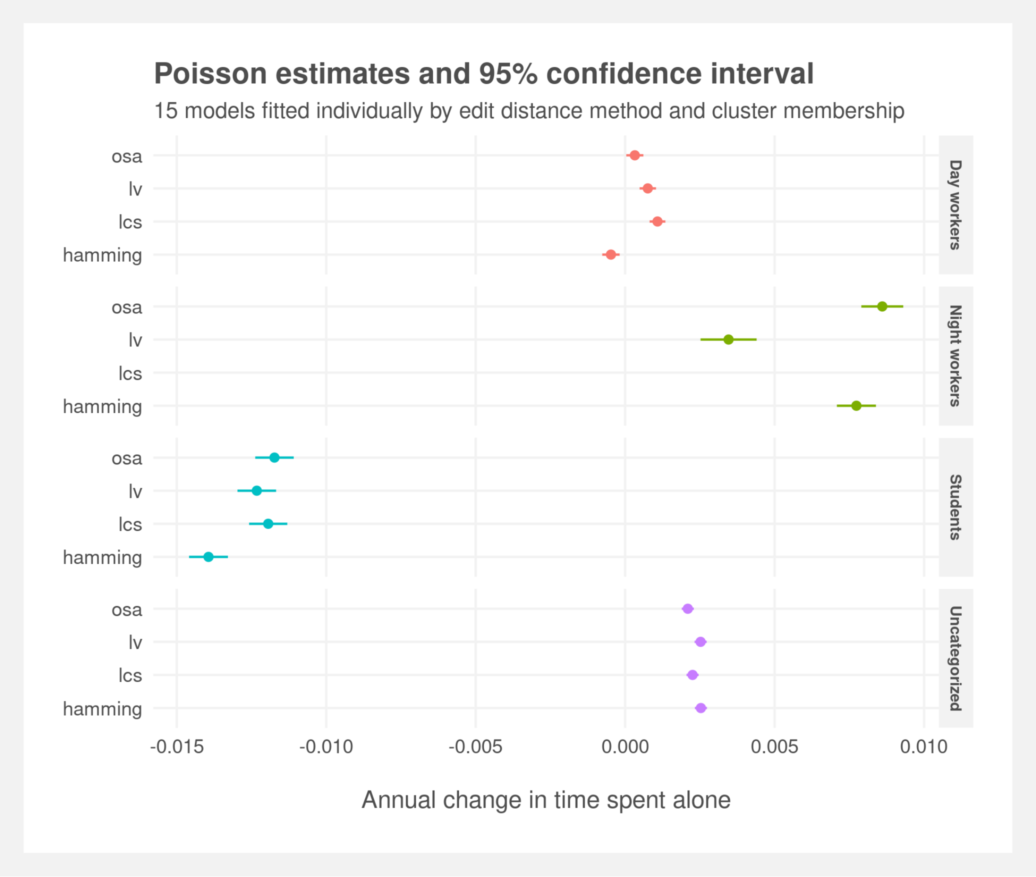

Figure 15: Poisson single-level effects

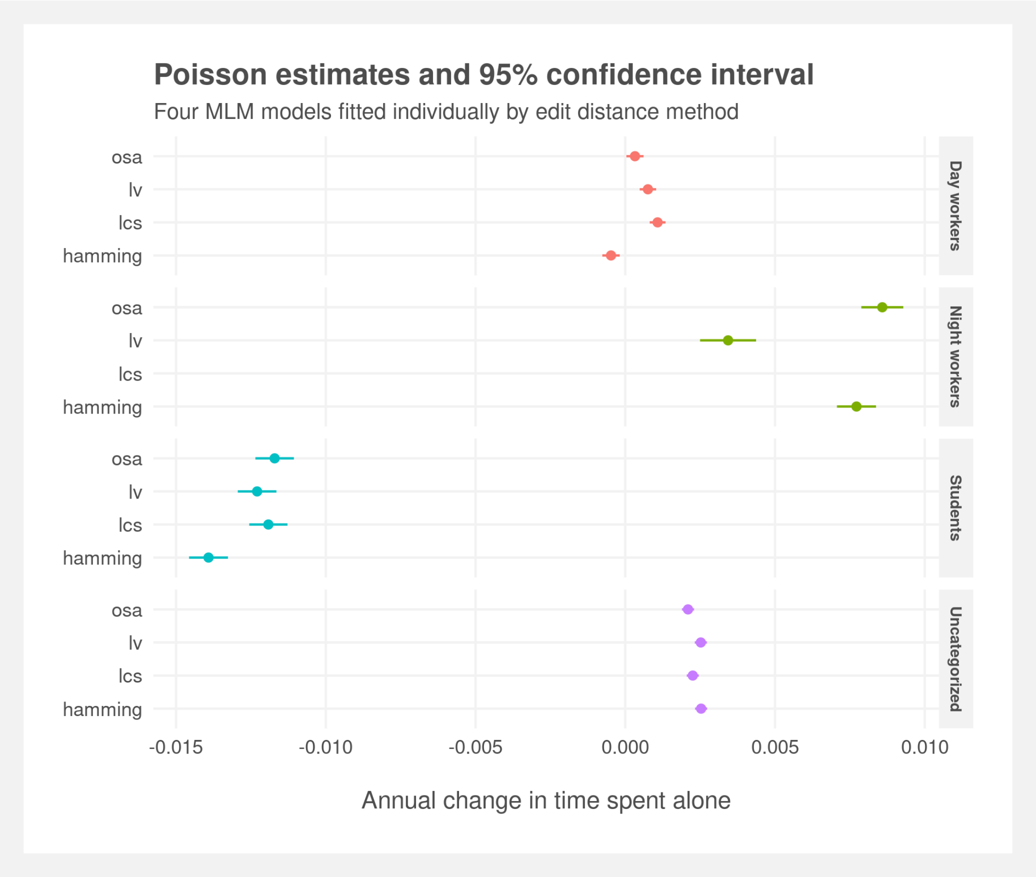

Figure 16: Poisson MLM effects

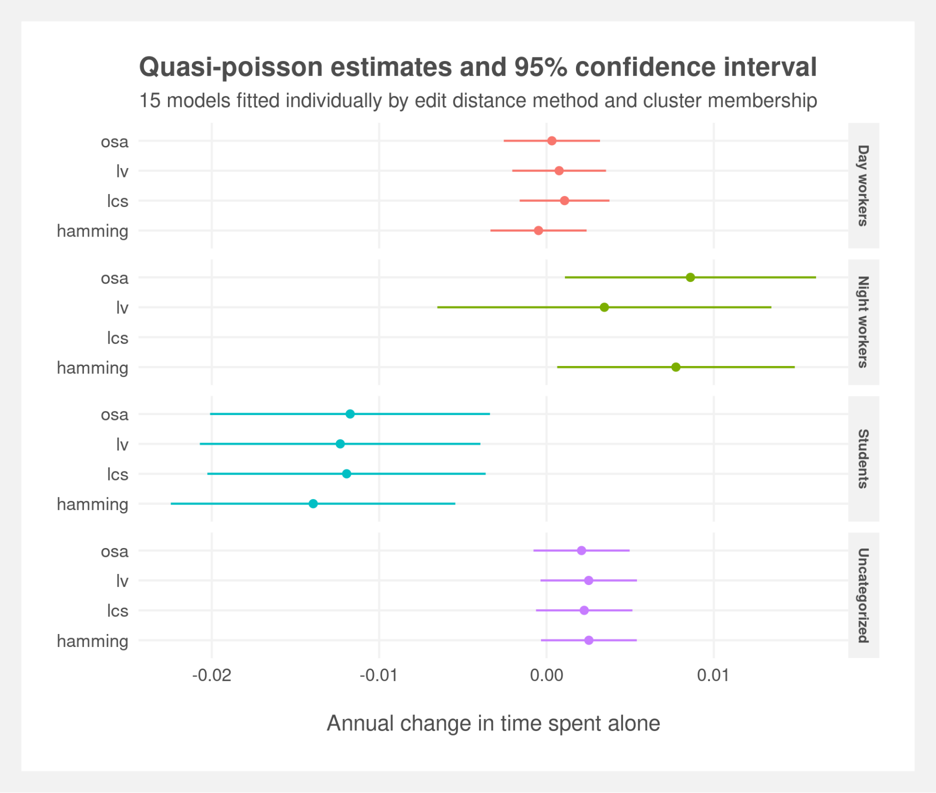

Figure 17: Quasi-Poisson estimates

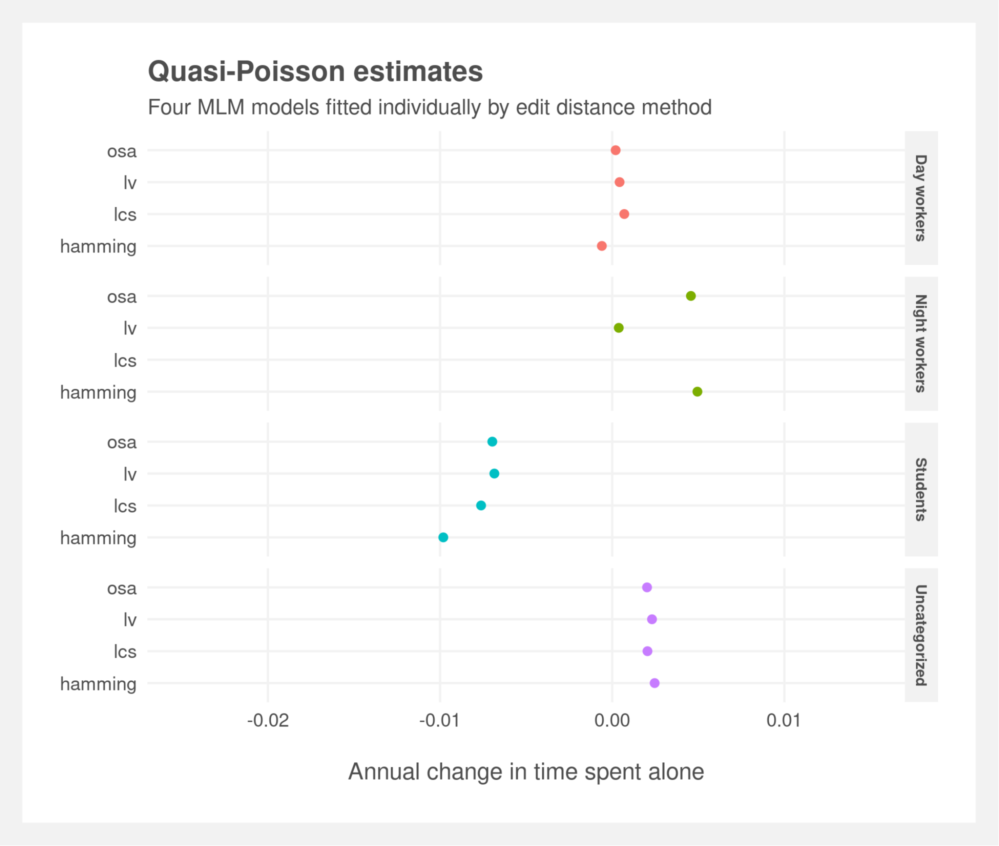

Figure 18: All edit distance measures: Multilevel quasi-Poisson estimates

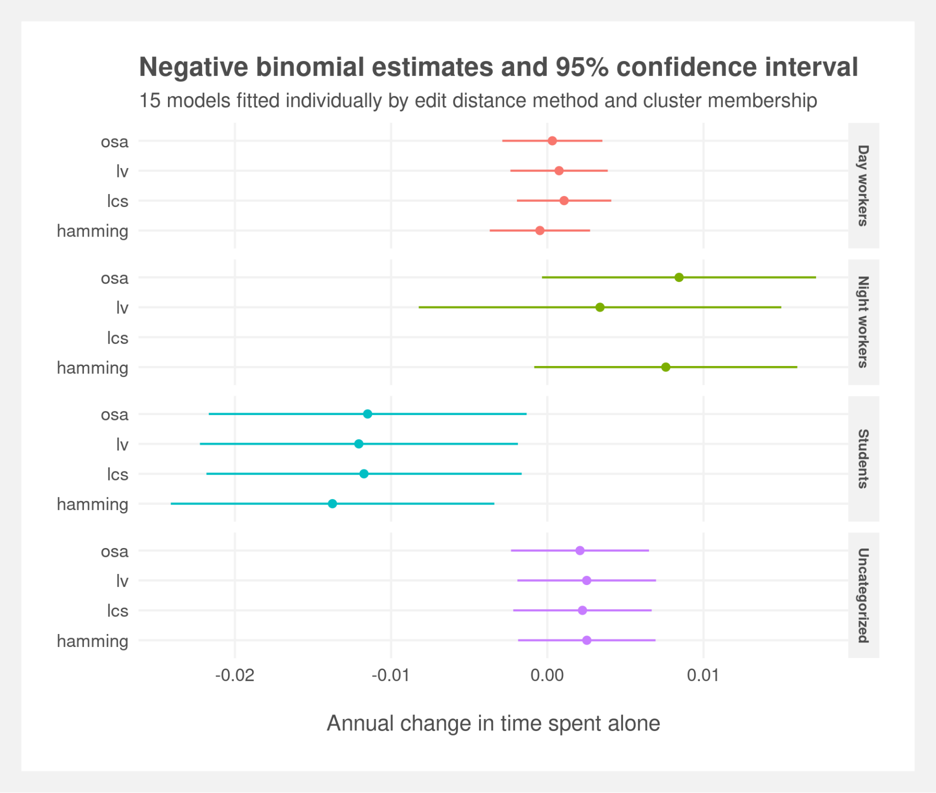

Figure 19: Negative binomial estimates

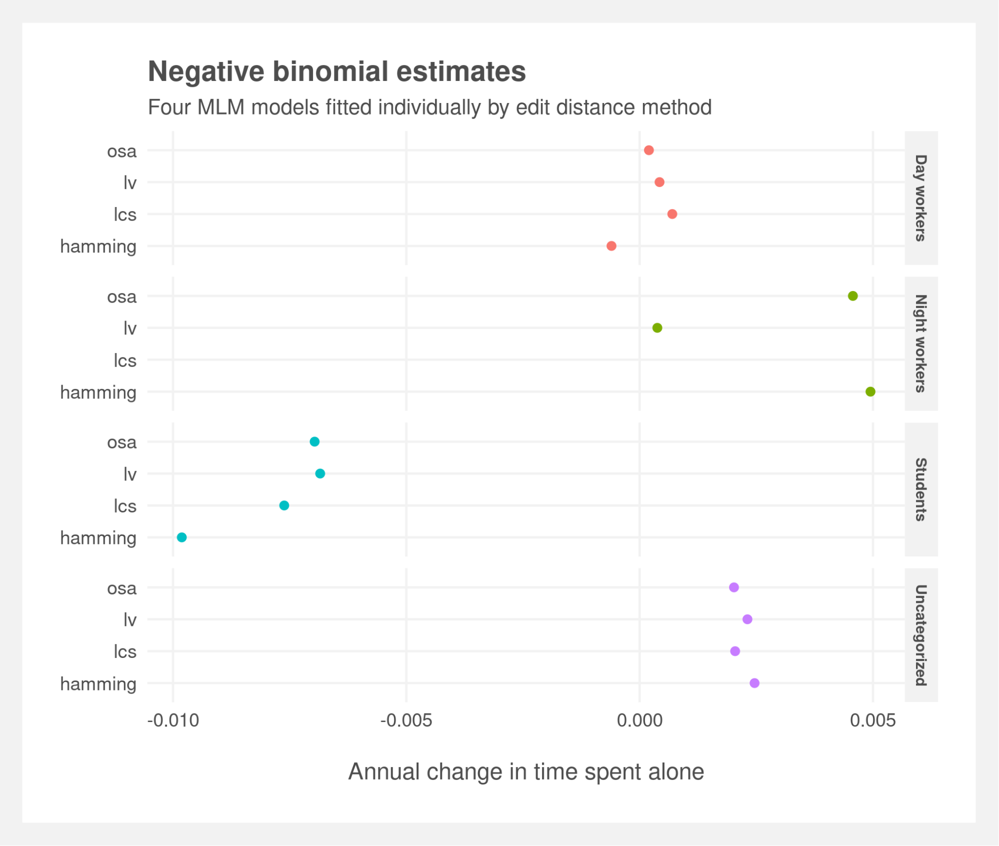

Figure 20: All edit distance measures: Multilevel negative binomial effects

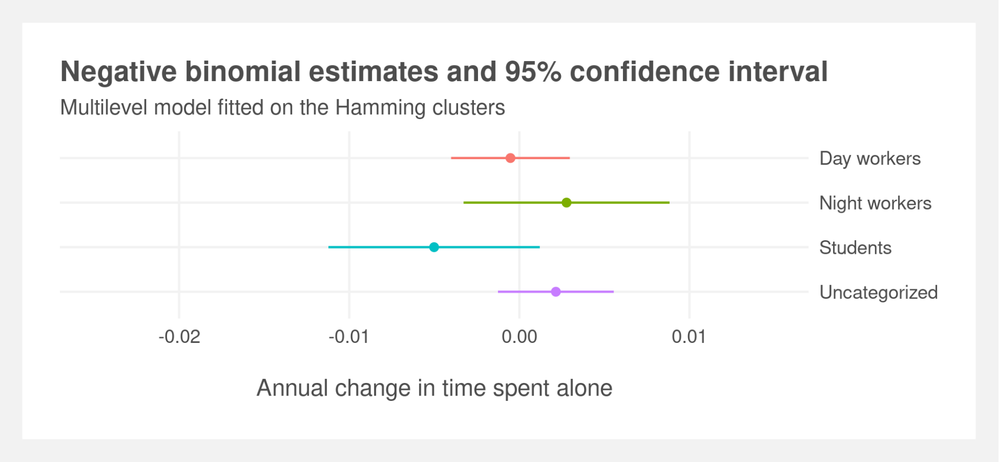

Figure 21: Hamming edit distance: Multilevel negative binomial effects

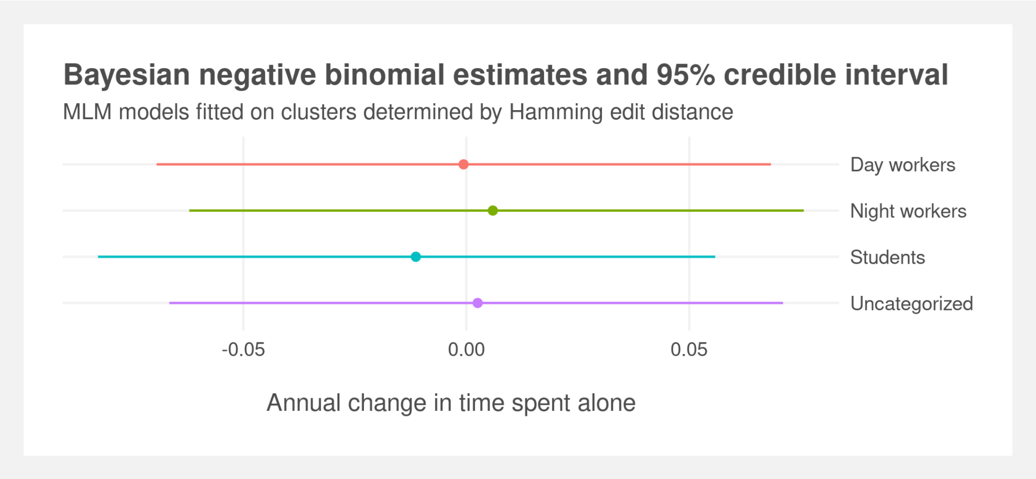

Figure 22: Bayesian negative binomial MLM effects