The AAAI-17 Workshop on
AI and Operations Research for Social Good
WS-17-01

Homelessness Service Provision: A Data Science Perspective

Yuan Gao, Sanmay Das, Patrick J. Fowler
Washington University in St. Louis
gao.yuan@wustl.edu, sanmay@wustl.edu, pjfowler@wustl.edu

Abstract

We study homeless service provision in the United States from a data science perspective, with the goal of informing homelessness prevention efforts. We use machine learning techniques to predict household reentry into a homeless system using an administrative dataset containing both demographic and service information. This data recorded all publicly funded services provided in a Midwestern US community from 2007 through 2014. We find that several techniques can provide useful lift in the prediction task, with random forests achieving an AUC around 0.7. Prediction improves significantly when conducted within calendar years, compared to across years, suggesting that changing dynamics drive repeated need for homeless services. We also analyze key service usage patterns that are associated with lower probabilities for reentry. Counterintuitively, individuals receiving the least intensive services provided through the homelessness system exhibit significantly lower likelihoods for further system involvement compared to individuals who received more intensive services, even after accounting for initial differences through propensity score and nearest neighbor matching. These result provide intriguing insights into homelessness service delivery that need to be further probed. In particular, it is unclear whether these less intensive services sustainably address housing needs, or whether, in contrast, frustration with inadequate services drives clients away from the homelessness system. Our results provide a proof-of-concept for how data science approaches can drive interesting, socially important research in the provision of public services.

Introduction

Homelessness represents an endemic and costly public health threat in nearly every community of the United States. On a single night in January 2014, there were an estimated 578,424 Americans experiencing homelessness, with the majority relying on emergency shelters or other temporary accommodations (Henry et al. 2014). Costs associated with homelessness significantly burden communities. Among a sample of chronically homeless individuals, Larimer et al. (2009) estimated total monthly costs of $4,066 per person between health care, criminal justice, and homeless services. Costly behavioral health care and foster care utilization are also elevated among the homeless, with emphasis on temporary shelters failing to reduce need, service use, and costs over time (Culhane, Park, and Metraux 2011).

The homeless system is primarily responsible for serving homeless individuals in most communities. Largely funded through the Department of Housing and Urban Development (HUD) Continuum of Care program, the system represents a network of local agencies partnering to provide a core set of resources that address basic housing needs. In 2014, this included 23,587 agencies in 416 communities across the country. Agencies develop and implement plans to rehouse the homeless that combine federally required universal elements with local discretion in service delivery. Fundamental services include access to emergency shelter as well as connection with short-term and permanent housing supports aimed to quickly rehouse people experiencing homelessness.

The homeless system aims to provide a flexible safety net for communities. However, the complexity involved in assessing and addressing homelessness has obfuscated the effectiveness of services. In particular, little rigorous research has examined how to accurately predict reentry into the homeless system after receipt of services, nor has research tested ways to optimize service delivery. Data science and machine learning applied to homeless service delivery provides a novel response to the public health threat.

At the same time, data from homeless systems provide novel and interesting challenges for machine learning. First, the prediction problems are themselves hard, as they often are in complex systems with human-generated data. For example, as our results in this paper show, it is difficult to achieve a performance over 0.7 in terms of the AUC measure for the problem of predicting whether an individual or family will re-enter the homeless system within two years of leaving. Second, it raises the classic question of how to estimate causal effects in the context of noisy data generated by human behavior. We take a first cut at doing so on a question of interest by using matching techniques. Specifically, we ask whether different types of interventions can lead to higher or lower probability of re-entry into the system within two years. Surprisingly, we find that the "lightest touch" intervention, which provides referrals and support rather than placement into a shelter, leads to significantly lower risk of re-entry. This holds true even when finding appropriate matched samples (using both propensity score and nearest neighbor matching) across observed characteristics, strengthening the case that this is a causal, rather than selection or confounding, effect. This raises the question of *why* this population is less likely to re-enter: Is it because the light touch intervention is successful, and reduces the occurrence of homelessness, or does it reflect a loss of faith in the system, so that this population doesn't even try to contact the system when they experience homelessness again in the future? Of course, our data cannot answer this question, since we don't gather any data from those who do not choose to contact the system again. However, our case study shows that taking a "data-first" approach to the problem can reveal interesting results that can, in turn, inform stakeholders in the community about important questions to follow up on.

Data

Data collection

A homeless management information system (HMIS) collected service records for all clients seeking federally funded housing assistance through the study period. The HMIS required collection of universal data elements that document client demographics, risks for housing problems, and all federally funded service interactions. Service providers entered client information in real time into a web-based software. A non-profit organization contracted by the homeless system hosted the software, as well as provided training and ongoing technical assistance for data quality.

Administrative records on homeless services were obtained through collaborative agreements with the governmental agencies responsible for collecting data, as well as the non-profit agency that hosted data entry software and warehoused information. The data were collected and stored according to federal mandates that included specification of universal data elements that captured information on client demographics and risk indicators, as well as dates and types of services.

The data that we use in this project was extracted from the original records of all services provided by 58 different homeless programs; individuals could be linked across programs by a unique and anonymous individual identifier. We aggregate data at the household level, merging individuals across files and removing duplicate entries. After this process, our data captures the beginning and end dates of all services across time, as well as household characteristics for both time-invariant (e.g., race, gender) and time-variant (e.g., income, public assistance receipt) indicators. The data provides an administrative timeline of services, including entry, transitions between service types, exits, and reentries across the study period.

Data features and outcomes of interest

We are primarily interested in the effectiveness of homelessness service provision. One measure of effectiveness is whether individuals or families experience repeat periods of homelessness, so we focus on reentry into the homeless service system as the primary dependent variable for our analyses. Operationally, we define reentry as contacting homeless services within 2 years following exit from the system over the study period. A two-year cutoff allows sufficient time to detect reentry and also coincides with federal guidelines for successful exits from the homeless system. Contact with homeless services was indicated by at least one request for housing assistance recorded in administrative data and deemed eligible. We use requests for services because waitlists frequently prohibited receipt of eligible services; thus, requests indicate *need*, while receipt represents *need plus availability*. We record an exit from the system by observing an end date of services that is separated by more than 1 day from the beginning of subsequent services. Transitions between homeless programming appear in the data when service dates overlap by 1 day or less, and thus, indicate continued service and not exit. Some households also receive multiple services during the same period; we include receipt of multiple services as a feature in our analyses.

Other indicators used to predict reentry come from universal data elements including household head age, monthly amount of income benefit, duration of first homeless service. These are typically measured continually. In addition, categorical indicators included the following of interest (the terms are relatively self explanatory): veteran status, disabling condition, prior residence, length of stay at prior residence, destination, reason of leaving, HUD chronic homeless, race, ethnicity, gender, has physical disability, received physical disability services, has developmental disability, received developmental disability services, has chronic health condition, received chronic health services, has HIV AIDs, received HIV AIDs services, has mental disability, received mental disability services, has substance abuse problem, received substance abuse services, domestic violence survivor, time of domestic violence occurred, income benefit type, source code of income benefit, overlapping visits, project type at entry of first shelter visit, project type at exit of first shelter visit, number of family members and number of children in the family.

Sample Characteristics

We focus on the population of households using homeless services from 2007 through 2014. This includes 9111 heads of households with adequate data of whom 3629 (39.8%) reentered the system multiple times. Restricting to families who exited before 2014, we are left with with 6836 households of whom 2928 (42.8%) reentered the system within two years of exit reentry.

Service Type	Service Type at Entry			Service Type at Initial Exit		
	% Dataset	% Reentry	% Change in Pr (Reentry)	% Dataset	% Reentry	% Change in Pr (Reentry)
Emergency Shelter	38.36	55.99	13.16	40.27	54.74	11.91
Rapid Rehousing	12.98	45.66	2.83	12.38	45.15	2.32
Temporary Housing	4.67	36.68	-6.15	27.40	38.87	-3.96
Permanent Supportive Housing	4.67	36.68	-6.15	4.18	37.06	-5.77
Homelessness Prevention	15.80	18.98	-23.85	15.77	19.02	-23.82

Table 1: Composition of the dataset by type of services provided, and the change in probability of reentry as a function of those services.

Analysis

We study two main questions: (1) Can we predict reentry into the homeless system based on demographic and other features? (2) How does placement into the lightest-weight intervention (homelessness prevention), as opposed to into a shelter or supportive housing, affect the probability of reentry into the system?

Predicting reentry

We tackle the problem of predicting reentry using a machine learning approach. We experiment with three different classification algorithms, namely decision trees, random forests, and logistic regression. Details of the training and how we measured accuracy are as follows

- Decision trees: We measure accuracy using 10-fold cross-validation. Each time, except for the 10% of the data left aside for testing, 80% of the remaining data was used to build a tree, and 20% for reduced-error pruning. We report area under the curve (AUC) score of the cross validation.
- Random forests: Each time, 1000 trees were used to build the classifier. The number of features randomly chosen as possibilities for split at each node was the square root of the total number of features. We report the out-of-bag (OOB) AUC score.
- Logistic regression with sum of the weights (L1) regularization: The regularization parameter was chosen by cross-validation. We report the AUC score of 10-fold cross-validation.

When comparing the methods, we find that random forests (AUC= .70) outperformed decision tree (AUC= .66) and logistic regression (AUC= .68) for this task. While the AUC numbers are not stunningly high, they do represent a meaningful improvement over base rates and are in line with other studies for difficult prediction tasks involving human behavior (e.g. (Das, Lavoie, and Magdon-Ismail 2013)).

Prediction accuracy over time

We next turn to the the question of how well the learning models generalize across time. We divided households into groups based on exit date of first homeless service: 2008-2009, 2010-2011, 2012-2013. We then trained random forests on each subgroup and applied these models to the other subgroups as well (similar to the technique used by Butaru et al. (2016) for credit risk analysis across time).

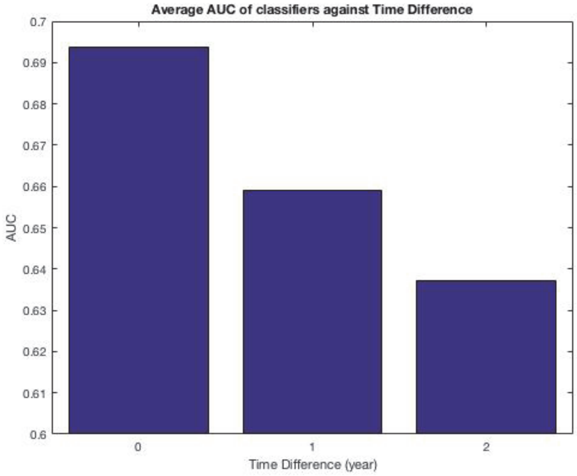


Figure 1: Average AUC of random forests as a function of the difference in time between the training and test data. 0 represents (OOB) AUC when testing on data from the same time period, while 1 and 2 are when the test data is separated from the training data by 1 or 2 years, respectively.

Our analysis shows that accuracy in classification decreases as a function of the temporal distance between the training and testing period, as illustrated in Figure 1. The AUC is highest with data trained from the same time period, consistently decreasing as the distance between training and test period increases. This shows that the factors driving repeat episodes of homelessness may well be dynamic rather than static, and any risk assessment tool should be responsive to such changes.

Feature analysis

Which features are most important in predicting reentry into the homeless system? We can analyze this by looking at the random forest importance scores. The top 15 scores, in rank order, are as follows:
destination after exit from the homeless system, duration of first homeless service spell, service type at initial exit from the system, monthly amount of income benefits, service type at entry into homeless services,

Homelessness Prevention vs:	Total pairs selected	Prevention reentered	% Prevention reentered	Non-Prevention reentered	% Non-Prevention reentered	Z-Score	p-value
Permanent Supportive Housing	161	25	15.5	59	36.6	-4.315	2.34E-05
Emergency Shelter	633	119	18.8	350	55.3	-13.44	0.00E+00
Temporary Housing	278	44	15.8	114	41.0	-6.5821	5.02E-11
Rapid Rehousing	381	63	16.5	170	44.6	-8.4131	2.92E-17

Table 2: Propensity score matching analysis of effects of homelessness prevention versus other interventions on homeless system reentry

Homelessness Prevention vs:	Total pairs selected	Prevention reentered	% Prevention reentered	Non-Prevention reentered	% Non-Prevention reentered	Z-Score	p-value
Permanent Supportive Housing	286	62	21.7	106	37.1	-4.04	7.39E-05
Emergency Shelter	1078	205	19.0	586	54.4	-17.025	1.16E-66
Temporary Housing	1078	205	19.0	462	42.9	-11.9742	2.11E-33
Rapid Rehousing	846	160	18.9	382	45.2	-11.567	2.56E-31

Table 3: Nearest neighbor matching analysis of effects of homelessness prevention versus other interventions on homeless system reentry

```
gender, race, age at entry, disabling
condition, type of income, length of
stay in residence before initial spell,
mental health problem, and chronic
health problem
```

The type of services received through the homeless system emerged as a key feature. Purely at the level of summary statistics, Table 1 shows that households receiving homeless prevention services (an intervention that typically provides networks of referrals, some help with payments, and case management) experienced substantial decreases in the probability of reentry, while those receiving emergency shelter services experienced increased likelihoods of reentry. Moreover, little change existed among households receiving more intensive and costly interventions of temporary housing and permanent supportive housing.

Accounting for observable differences

Our initial analysis suggests that households allocated to homelessness prevention, rather than an alternative, heavier-weight intervention, are less likely to seek homelessness services. However, this could be a selection effect (or a confounding effect) – perhaps those being allocated to the lighter-weight intervention are those who are less at risk in the first place?

To explore these differences further, we used two different (one-to-one) matching methods (Stuart 2010) to account for observable differences between groups that may explain outcomes. Propensity score matching adjusts for potential differences in being assigned to the different service types, while nearest neighbor matching makes comparisons based on the most similar individuals regardless of service to account for any potential bias in referrals. We performed various sensitivity analyses to ensure the adequacy of matches, and then tested average differences in the probability of reentry into the homeless system between service types adjusted for pre-existing differences between groups. The two groups in each analysis were (1) households who received homeless prevention and (2) similar households referred for each of the other types of homeless services.

Comparisons using both propensity score and nearest neighbor matching strongly support the conclusion that those households who receive homelessness prevention services are significantly less likely to reenter the homeless system within two years, and that the effect size is large. Table 2 shows that after accounting for the propensity of being referred for homelessness prevention, those who received prevention services were less likely to return to the system within two years. Likewise, Table 3 suggests that substantial differences persist even after accounting for potential non-random processes involved in receiving different service types.

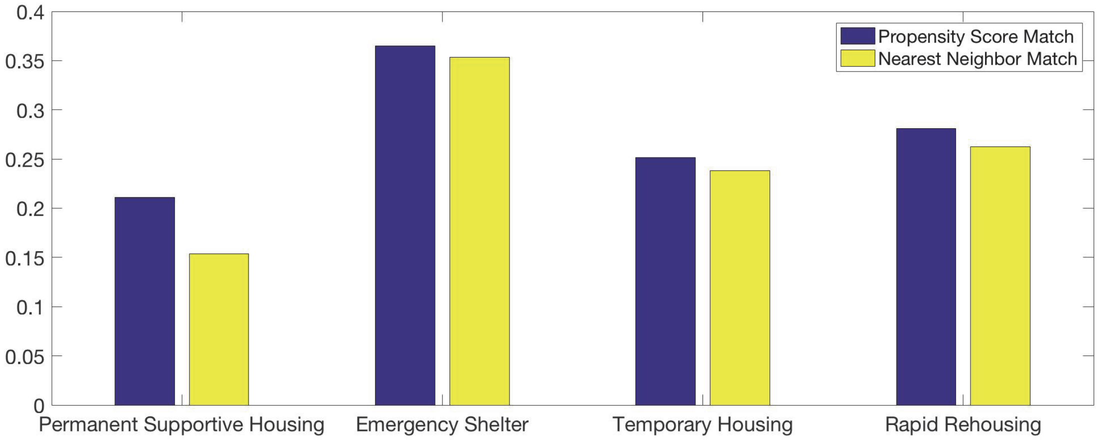
Figure 2 shows a comparison of the two matching methods. It shows the percentage point differences of reentry between each of the four shelter types and Homelessness Prevention, measured by the two different types of matching methods. It shows that, although there are slight percentage differences, the patterns are similar in both matching results.

Conclusion

This preliminary study illustrates the potential role as well as limits of data science applied to efforts to prevent homelessness. Machine learning using all universally available data from service provision improves the prediction of reentry substantially beyond base rates; however, a nontrivial amount of uncertainty remains in prediction. Additional research needs to understand the extent to which these methods improve beyond current practice. However, there is potential for such tools to become useful for identifying at-risk populations and intervening appropriately.

Our findings also yield intriguing counterintuitive patterns associated with homeless services receipt. Homelessness prevention provides substantial decreases in reentry into the homeless system despite representing the least intensive and cheapest service option. Although it is tempting to conclude that homelessness prevention represents the best option to

Figure 2: Percentage point differences in reentry probability as assessed by both matching methods



end homelessness, it remains unclear whether failure to return to the homeless system truly indicates reduced housing risk. Alternatively, homeless prevention could unintentionally disenfranchise households by woefully failing to meet needs; the lack of effectiveness of these services may cause individuals to avoid the homeless system. We cannot rule out the plausibility of such a hypothesis using the available data, but it does raise a question that can be further engaged with stakeholders. Future research with homeless individuals and service providers will further explore ways to optimize delivery of homeless services.

Acknowledgments

We thank our community partners who provided data and information on the homeless management information system.

References

Butaru, F.; Chen, Q.; Clark, B. J.; Das, S.; Lo, A. W.; and Siddique, A. R. 2016. Risk and risk management in the credit card industry. *Journal of Banking and Finance* 72:218–239.

Culhane, D. P.; Park, J. M.; and Metraux, S. 2011. The patterns and costs of services use among homeless families. *Journal of Community Psychology* 39(7):815–825.

Das, S.; Lavoie, A.; and Magdon-Ismail, M. 2013. Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. In *Proceedings of the ACM Conference on Information and Knowledge Management*, 1097–1106.

Henry, M.; Cortes, A.; Shivji, A.; and Buck, K. 2014. The 2014 Annual Homelessness Assessment Report (AHAR) to Congress: Part 1 Point-in-Time Estimates of Homelessness. The Department of Housing and Urban Development.

Larimer, M. E.; Malone, D. K.; Garner, M. D.; Atkins, D. C.; Burlingham, B.; Lonczak, H. S.; Tanzer, K.; Ginzler, J.; Clifasefi, S. L.; Hobson, W. G.; et al. 2009. Health care and public service use and costs before and after provision of housing for chronically homeless persons with severe alcohol problems. *Journal of the American Medical Association* 301(13):1349–1357.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25(1):1–21.