ICOSSAR 2021
The 13th International Conference on Structural Safety and Reliability (ICOSSAR 2021), June 21-25, 2021, Shanghai, P.R. China
J. Li, Pol D. Spanos, J.B. Chen & Y.B. Peng (Eds)

Logistic Models Linking Household Recovery Capacity to Demographic Characteristics

Rodrigo Costa [a], Chenbo Wang [b], and Jack W. Baker [c]

[a] *Department of Civil and Environmental Engineering, Stanford University, E-mail: rccosta@stanford.edu.*
[b] *Department of Civil and Environmental Engineering, Stanford University, E-mail: wangcb@stanford.edu.*
[c] *Department of Civil and Environmental Engineering, Stanford University, E-mail: bakerjw@stanford.edu.*

American Community Survey data are used to build logistic regression models that predict the capacity of owner households to finance the post-earthquake repair of their homes. Having high income and a paid mortgage are used as proxies for the ease of access to financing. We find that households with limited knowledge of English and those with elderly members are the least likely to have a high income. Households with high income and those who have recently moved to their current homes are the most likely to have a mortgage. An example application considering post-earthquake housing recovery in San Francisco is presented. The example demonstrates a strong disparity in the recovery capacity of households based on their demographics. The developed models have important implications for post-earthquake housing recovery as they help to identify the demographics that make households less capable of repairing their homes after an earthquake.

1 Introduction

Due to the rare nature of earthquakes, our ability to anticipate the factors that will dictate post-earthquake housing recovery in future disasters is limited. For this reason, it is generally accepted that predictive models of housing recovery can provide valuable insights for recovery planning. Simulation models have been used to investigate housing recovery after hypothetical earthquakes in Jerusalem (Grinberger and Felsenstein, 2016), the south Napa (Kang et al., 2018), Nepal (Longman and Miles, 2019), Vancouver (Costa et al., 2020), the San Francisco Bay Area (Markhvida et al., 2020), and the Lombok region in Indonesia. (Alisjahbana and Kiremidjian, 2020). These simulations demonstrate that low-income households are affected disproportionately and are less capable of repairing their homes. The impact of income on recovery capacity is relatively intuitive, e.g., wealthier households are more likely to have the means to repair their homes. From a housing recovery modeling perspective, disparities in the recovery capacity of different income groups can be simulated via their ease to access financing for repairs.

However, to inform the development of housing recovery policies, it is important to understand the correlation between recovery capacity and a wider range of household demographics. Studies of previous disasters have demonstrated that housing recovery is made uneven by demographics such as home ownership (Wu, 2004; Kamel and Loukaitou-Sideris, 2004; Mayer et al., 2020), race (Bullard and Wright, 2009; Fussell et al., 2010; Peacock et al., 2014), immigra-

tion status and linguist barriers (Kamel and Loukaitou-Sideris, 2004; Loukaitou-Sideris and Kamel, 2004), education (Burton, 2015; Nejat et al., 2019), family structure (Nejat and Ghosh, 2016; Nejat, 2018), age (Ngo, 2001; Henderson et al., 2010), gender (Nejat et al., 2018) household size (Sadri et al., 2018), among others. Unlike for income, there are no simple heuristics to describe the effect of demographics such as immigration status, family structure, or race on housing recovery capacity. Empirical studies have demonstrated that certain groups are less capable of repairing their homes. However, the mechanisms that lead to differential recovery capacity in previous disasters may not be transferable across space and time. Thus, simulating the disparity between households due to their demographics is challenging.

Regression models have been employed to investigate the demographics that limited the housing recovery capacity of households in past events. A recent study by Nejat et al. (2019) used survey data to build a spatial logistic regression model to predict the recovery decisions made by households in Staten Island, New York, in the aftermath of Hurricane Sandy. However, a gap still exists when the goal is to use simulations to predict the demographics that are are expected to limit the capacity of households to repair their homes.

To address this gap, this paper develops logistic regression models that predict household income and mortgage status, which in turn are used in housing recovery simulations as proxies for the ability of the households to finance home repairs. With this approach, a wider range of household demographics associated with the capacity to quickly recover are identified. The methodology relies on publicly available information from the American Community Survey. Thus, it can be applied to any community for which data are directly available from the survey. This can inform targeted mitigation policies, as well as assist in evaluating their potential benefits.

2 Housing Recovery Simulation

In this section we discuss how the logistic regression models developed in this study can be incorporated into housing recovery simulations. Figure 1 shows the proposed algorithm for simulating housing recovery, which is similar to other existing models (Sutley et al., 2017; Burton et al., 2018; Costa et al., 2020). Here, the focus is on owner-occupied single-family buildings because the recovery of these buildings is dictated by the households that occupy them. This study assumes that all households would be willing to repair their buildings after an earthquake. The algorithm in Figure 1 considers the recovery of a portfolio of N_B damaged buildings is of interest. The time needed to repair building i, $T_{r,i}$, is estimated using a HAZUS-like approach (FEMA, 2015). However, before repairs can start, financing, materials, and skilled workers must be procured. The algorithm considers four sequential decisions, represented as rhombuses. Decision (1) guarantees that the state of each building is evaluated on each time step. Decision (2) checks if building i has not yet been fully repaired. Next, the availability of funds is checked in Decision (3). If financing is obtained, Decision (4) checks if materials and workers are available for repairs. If Decision (4) returns 'No', the household enters a competition for these resources against other households that have already obtained financing. If Decision (4) returns 'Yes,' this means the building can advance repairs, and the duration of the current time step, Δt, is deducted from the total time need to repair the building, $T_{r,i}$. The building is considered repaired when $T_{r,i}$ =0. The algorithm is evaluated repeatedly to simulate the progress along the timeline of recovery.

Decision (3) in Figure 1 is the main focus of this study. Funds from insurance, private loans, and public loans are considered. The REDi guidelines (Almufti and Willford, 2013) provide estimates of the times needed to obtain financing from these three sources. Insurance payments are the fastest alterna-

ICOSSAR 2021
The 13th International Conference on Structural Safety and Reliability (ICOSSAR 2021), June 21-25, 2021, Shanghai, P.R. China
J. Li, Pol D. Spanos, J.B. Chen & Y.B. Peng (Eds)

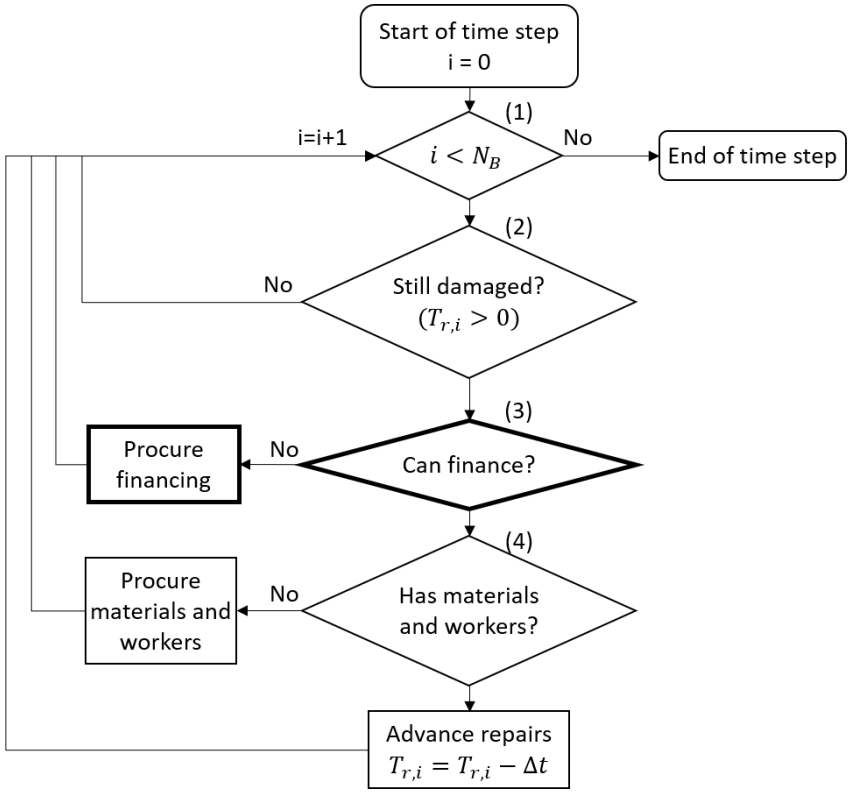

Figure 1. Algorithm for housing recovery simulation illustrating the decisions made by households on damaged buildings on one time step of the simulation. The highlighted elements indicate the focus of this study.

tive, followed by private loans. Thus, insured households obtain financing and procure materials and workers sooner than those that rely on loans. The number of construction workers in the community is limited and allocated on a first-come-first-serve basis. Thus, the households that are the slowest at obtaining financing may need to wait for repairs to be completed on other buildings before having access to certain resources.

Figure 2 shows the semi-heuristic algorithm for repair financing used in this study. Earthquake insurance take-up rates in California are low, about 10% (Marshall et al., 2018). The algorithm assumes that only, but not all, high-income households have insurance. A deductible equal to 15% of the value of the building is assumed (Marshall et al., 2018). If the deductible is below the annual household income, the household is assumed to be able to pay it out-of-pocket. Low-income and moderate-income homeowners, those not insured, and those insured who cannot pay the deductible out-of-pocket seek a loan. It is assumed that homeowners with paid mortgages can obtain all funds needed for repairs from a private loan. Furthermore, homeowners with mortgages would depend on public loans. Homeowners are assumed to need to obtain all financing before repairs can start.

The algorithm in Figure 2 relates income and mortgage status to the mechanism used by a household to finance repairs. This paper develops logistic regression models that associate a household's demographic characteristics to the probability that it has a high income or a mortgage. These logistic regression models, discussed in the following, allow for the recovery capacity of specific socioeconomic groups to be partially captured by the simulation model.

3 Logistic Regression Models

The American Community Survey annually collects data on the population demographics and the housing stock of major US cities and distributes them as Public Use Microdata

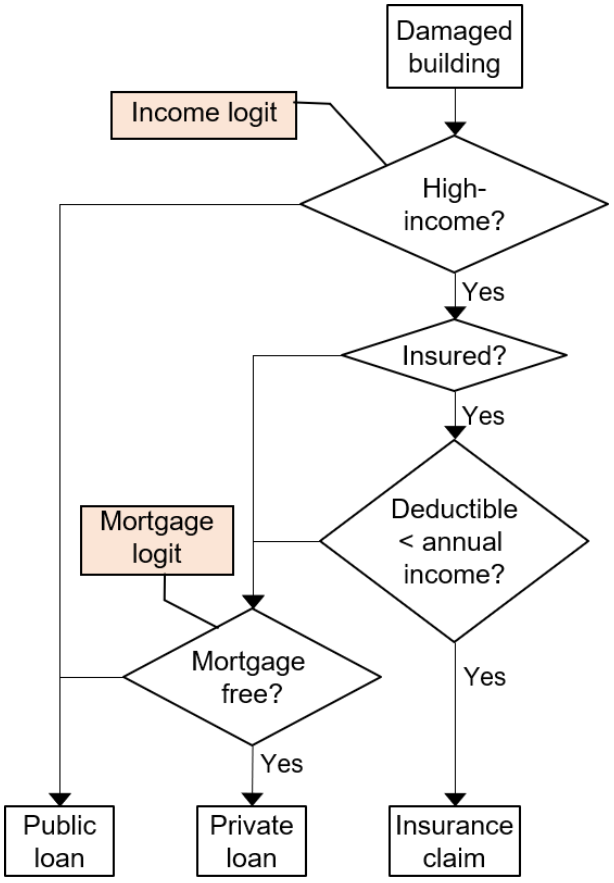

Figure 2. Financing algorithm for housing reconstruction indicating the logistic regression models for income and mortgage status.

Samples (PUMS). The PUMS files are a set of records from individual households available in three formats: 1-year counts, 3-year counts, or 5-year counts. As per the US Census Bureau, the 5-year counts are the best data for analyses not significantly affected by recency. For this reason, this study employs the 2014-2018 PUMS.

The American Community Survey PUMS are used in this study to build logistic regression models to predict if a household has a high income or a mortgage. This study focuses on single-family owner-occupied homes in the city of San Francisco, for which there are 4878 survey responses available. Table 1 lists the American Community Survey demographics selected as possible predictors of household income and mortgage status. These demographics are selected because they have been demonstrated to be correlated with social vulnerability. The variable 'Moved in after 2010' is aimed at capturing if the household is a long-term resident of San Francisco or not. Two demographics in Table 1 are marked with (*) to indicate they are estimated via proxies. The households which indicated not to have any member who fluently speaks English were categorized as having a limited knowledge of English. The households which indicated that the language spoken among household members is Spanish are categorized as having a Hispanic background.

The demographics in Table 1 are collected in a vector of n predictor variables $\boldsymbol{X} = 1, X_1, \ldots, X_n$, and used to fit logistic regression models. Boolean yes/no data are mapped into 1/0 variables. Categorical data with c categories are mapped into c-1 dummy variables. The probability that a household has a high income (or a mortgage) given its demographics, P(HI=yes$|\boldsymbol{X}$), is estimated as

$$\mathrm{P(HI{=}yes}|\mathbf{X}) = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{X})}{1+\exp(\boldsymbol{\beta}^T \boldsymbol{X})} \qquad (1)$$

where $\boldsymbol{\beta} = \beta_0, \ldots, \beta_n$ are the coefficients of the model, which are fitted using the Least Absolute Shrinkage and Selection Operator (LASSO). The LASSO imposes a penalty on the size of the coefficients $\boldsymbol{\beta}$, forcing predictors with low predictive power to be selected out of the final model (Tibshirani, 1996). Due to the regularization, certain coefficients are shrunk to zero leading to models with fewer predictors. Models with fewer predictors have been shown to be less prone to overfitting and to yield better predictions when applied to new data (Agresti, 2003).

The first logistic regression model estimates the probability that a household has a high income, i.e., P(HI=yes$|\boldsymbol{X}$). High income is defined as income higher than 120% of the median area income, about $115,000 for San Francisco. Table 2 shows the significant variables identified by the LASSO and their coefficients. The area under the receiver operating characteristic curve, i.e., AUC, for the model in Table 2 is 0.79. The results show that

Table 1. Household demographics available from the American Community Survey. Demographics marked with * are estimated from proxies.

Demographic	Categories
Building type	Single-family or Multi-family
Building year built	$\leq$ 1941, 1940 to 1975, or $>$ 1975
Building value	Real number
Household size	Integer number
Moved in after 2010	Yes or No
Married couple	Yes or No
Children in household	Yes or No
Elderly in household	Yes or No
Limited knowledge of English*	Yes or No
Hispanic background*	Yes or No

households with limited knowledge of English are the least likely to have a high income as denoted by the high negative value of the log-odds estimate for this predictor. Households comprised of married couples, with no elderly, and those who have recently moved into the city are more likely to have high incomes as well. In terms of size, the larger the household, the more likely it is to have high income.

Table 2. Regression coefficients for P(HI=yes$|\boldsymbol{X}$). All predictors are significant at the $p < 0.001$ level.

Predictor	Estimate
Intercept	-1.49
Limited knowledge of English	-1.76
Elderly in household	-0.64
Moved in after 2010	0.57
Married couple	0.87
Building value [M$]	0.38
Household size	0.35

The second model estimates the probability that a household has a mortgage on their home, P(mortgage=yes$|\boldsymbol{X}$). Table 3 shows the demographics identified as significant predictors in this model. Households with limited knowledge of English and those with elderly members are the least likely to have a mortgage. Households with high income and who have recently moved into their homes are the most likely to have a mortgage. From the American Community Survey data it is observed that households with limited knowledge of English tend to occupy lower-valued buildings, have elderly members, and not to be recently movers. Recently mover households tend to be high-income, younger, married couples with children who occupied highly-valued buildings. This helps explains the counter-intuitive fact that households with limited knowledge of English tend not to have a high-income and yet have paid mortgages. The AUC for the model in Table 3 is 0.73.

Table 3. Regression coefficients for P(mortgage=yes$|\boldsymbol{X}$). All predictors are significant at the $p < 0.001$ level.

Predictor	Estimate
Intercept	0.22
Limited knowledge of English	-0.44
Elderly in household	-0.69
Moved in after 2010	0.73
Children in household	0.71
High-income	0.74

4 Case Study and Results

This illustrative case study simulates the housing recovery of single-family owner-occupied homes in San Francisco after an

M_w=7.9 earthquake on the San Andreas fault. All buildings are assumed to be light-frame wood buildings, i.e., 'W1' as per the HAZUS classification (FEMA, 2015). Each household and building in the simulation is given a set of demographic characteristics (Table 1) sampled from the population distribution. San Francisco is comprised of 194 Census tracts. The attributes of the households are sampled from Census data for the Census tract it belongs to. For example, if Census data indicate that 30% of all households in a Census tract have a low income, 30% of all households are randomly assigned as low income. The use of Census tract data to instantiate the household demographics allows for the spatial correlations between demographics, e.g. having a limited knowledge of English and being a recent mover, to be partially captured. Although this is outside of the scope of this study, explicitly accounting for the correlation between these demographics would theoretically lead to more refined results. Once the building and household demographics are attributed, whether the household has a high income or a mortgage is then simulated according to the probabilities based on Eq. 1 and Tables 2 and 3, using the household characteristics as inputs.

Fragility curves are used to translate the ground motion intensities estimated at the centroid of the census tracts into building damage. The algorithm in Figure 1 is then used to simulate housing recovery over time. This case study investigates recovery for a period of five years after the earthquake. During this period, it is assumed that at most 10% of the buildings that were initially damaged by the earthquake can be under repairs at the same time. The 10% ceiling is arbitrarily imposed in order to create scarcity of resources.

Figure 3 shows the housing recovery curves for selected demographics. In each figure panel the results are deaggregated based on the categories of the indicated demographic. The solid lines in these panels are the number of buildings which are damaged and have not yet been repaired at a given time, indicated on the left-hand side ordinate axis. For example, the black solid line on the top-left panel indicates that immediately after the earthquake nearly 17,000 damaged buildings are owned by high-income households. There are more damaged buildings occupied by high-income households because there are more high-income households in general, i.e., filtering is not accounted for. The slopes of the solid lines indicate the speed at which recovery is progressing. If the rate of change in these slopes is calculated and plotted against time since the event, the dotted lines in each panel are obtained. Note that the ordinate axes for the dotted lines are plotted on the right-hand side of each panel. The dotted lines are not dependent on the number of buildings damaged in each category, providing a normalized metric of performance of housing recovery.

The dotted lines show that the disparity in the speed of housing recovery is highest between the high-income and non-high-income groups. The disparity between income groups is highest during the first year but is noticeable throughout the period investigated. Limited knowledge of English has the second most noticeable effect on the speed of housing recovery. This is in accordance with the logistic regression model for high income, which indicated that households with limited knowledge of English are the least likely to have high income. Similarly, households comprised of married couples fare better because this is a good predictor of higher income. The effect of the demographics in the three panels in the bottom is less noticeable. It is also noted that unlike the effect of high income, the effect of the remaining demographics is less pronounced beyond the first year. This is due to these demographics not being perfect predictors of having high income.

The results in Figure 3 demonstrate that the proposed methodology can at least partially capture disparities in the recovery capacity of households with different demographics. These disparities are captured without imposing strict assumptions of the effects that a de-

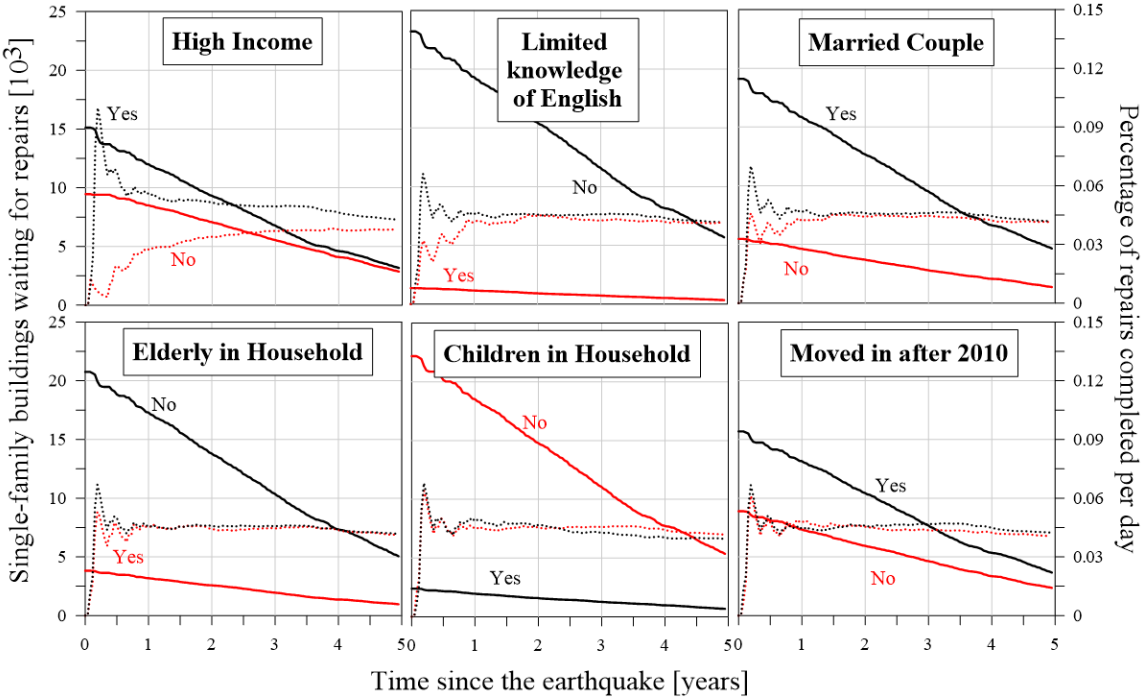

Figure 3. Results by selected demographics. The solid lines are the number of homes waiting for repairs, indicate on the left-hand side ordinate axis. The dotted lines indicate the speed of recovery, measures as the % of repairs completed per day since the earthquake.

mographic, e.g., Hispanic background, has on recovery capacity. Household demographics are simply correlated to income and mortgage status. If this correlation was not accounted for, i.e., Hispanic background, income, and mortgage status of households were assigned randomly, the dotted lines in Figure 3 would be parallel.

5 Conclusions

This paper introduces a methodology to associate a wide range of household demographics to their ability to finance the post-earthquake repairs of their homes. Two logistic regression models fitted from data from American Community Survey of San Francisco are developed. These models are then used to estimate how quickly a household can obtain financing based on its demographics, family structure, and income. The speed at which households obtain financing is used in housing recovery simulations to determine when these households can start procuring resources for reconstruction. Thus, the proposed methodology can capture the influence of household demographics on their recovery capacity. Quantifying potential difficulty in recovery provides insights into disaster recovery planning, and helps identify effective actions to improve housing recovery. The regression models in this paper can be used to improve the results produced by existing housing recovery models. Furthermore, the models are fitted using publicly available data which are collected for several communities in the US. It is envisioned that the proposed methodology can be used to study recovery capacity of different socioeconomic groups in any of these communities.

6 Acknowledgments

This work was supported in part by the Stanford Urban Resilience Initiative and the Stanford UPS Endowment Fund.

7 References

Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

Alisjahbana, I. and Kiremidjian, A. (2020). Modeling housing recovery after the 2018 lombok earthquakes using a stochastic queuing model. *Earthquake Spectra*, page 8755293020970972.

Almufti, I. and Willford, M. (2013). REDi Rating System: Resilience Based Earthquake Design Initiative for the Next Generation of Buildings. Version 1.0. Technical report, Arup.

Bullard, R. D. and Wright, B. (2009). *Race, place, and environmental justice after Hurricane Katrina: Struggles to reclaim, rebuild, and revitalize New Orleans and the Gulf Coast*. PERSEUS BOOKS.

Burton, C. G. (2015). A validation of metrics for community resilience to natural hazards and disasters using the recovery from hurricane katrina as a case study. *Annals of the Association of American Geographers*, 105(1):67–86.

Burton, H. V., Miles, S. B., and Kang, H. (2018). Integrating performance-based engineering and urban simulation to model post-earthquake housing recovery. *Earthquake Spectra*, 34(4):1763–1785.

Costa, R., Haukaas, T., and Chang, S. E. (2020). Agent-based model for post-earthquake housing recovery. *Earthquake Spectra (in press)*.

FEMA (2015). HazusMH 2.1: Technical Manual. Technical report, Federal Emergency Management Agency.

Fussell, E., Sastry, N., and VanLandingham, M. (2010). Race, socioeconomic status, and return migration to new orleans after hurricane katrina. *Population and environment*, 31(1-3):20–42.

Grinberger, A. Y. and Felsenstein, D. (2016). Dynamic agent based simulation of welfare effects of urban disasters. *Computers, Environment and Urban Systems*, 59:129–141.

Henderson, T. L., Roberto, K. A., and Kamo, Y. (2010). Older adults responses to hurricane katrina: Daily hassles and coping strategies. *Journal of Applied Gerontology*, 29(1):48–69.

Kamel, N. M. and Loukaitou-Sideris, A. (2004). Residential assistance and recovery following the Northridge earthquake. *Urban Studies*, 41(3):533–562.

Kang, H., Burton, H. V., and Miao, H. (2018). Replicating the recovery following the 2014 south napa earthquake using stochastic process models. *Earthquake Spectra*, 34(3):1247–1266.

Longman, M. and Miles, S. B. (2019). Using discrete event simulation to build a housing recovery simulation model for the 2015 nepal earthquake. *International Journal of Disaster Risk Reduction*, 35:101075.

Loukaitou-Sideris, A. and Kamel, N. M. (2004). Residential Recovery from the Northridge Earthquake: An Evaluation of Federal Assistance Programs. Technical report, CALIFORNIA POLICY RESEARCH CENTER. [Online; accessed 15-May-2019].

Markhvida, M., Walsh, B., Hallegatte, S., and Baker, J. (2020). Quantification of disaster impacts through household well-being losses. *Nature Sustainability*, pages 1–10.

Marshall, D. et al. (2018). An overview of the california earthquake authority. *Risk Management and Insurance Review*, 21(1):73–116.

Mayer, J., Moradi, S., Nejat, A., Ghosh, S., Cong, Z., and Liang, D. (2020). Drivers of post-disaster relocations: The case of moore and hattiesburg tornados. *International Journal of Disaster Risk Reduction*, page 101643.

Nejat, A. (2018). Perceived neighborhood boundaries: A missing link in modeling post-disaster housing recovery. *International journal of disaster risk reduction*, 28:225–236.

Nejat, A., Brokopp Binder, S., Greer, A., and Jamali, M. (2018). Demographics and the dynamics of recovery: A latent class analysis of disaster recovery priorities after the 2013 moore, oklahoma tornado. *International Journal of Mass Emergencies & Disasters*, 36(1).

Nejat, A. and Ghosh, S. (2016). LASSO Model of Postdisaster Housing Recovery: Case Study of Hurricane Sandy. *Natural Hazards Review*, 17(3):1–13.

Nejat, A., Moradi, S., and Ghosh, S. (2019). Anchors of social network awareness index: A key to modeling postdisaster housing recovery. *Journal of Infrastructure Systems*, 25(2):04019004.

Ngo, E. B. (2001). When disasters and age collide: Reviewing vulnerability of the elderly. *Natural Hazards Review*, 2(2):80–89.

Peacock, W. G., Van Zandt, S., Zhang, Y., and Highfield, W. E. (2014). Inequities in long-term housing recovery after disasters. *Journal of the American Planning Association*, 80(4):356–371.

Sadri, A. M., Ukkusuri, S. V., Lee, S., Clawson, R., Aldrich, D., Nelson, M. S., Seipel, J., and Kelly, D. (2018). The role of social capital, personal networks, and emergency responders in post-disaster recovery and resilience: a study of rural communities in indiana. *Natural hazards*, 90(3):1377–1406.

Sutley, E. J., Peek, L., and van de Lindt, J. W. (2017). Community-Level Framework for Seismic Resilience. I: Coupling Socioeconomic Characteristics and Engineering Building Systems. *Natural Hazards Review*, 18(3):4016014.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wu, J. Y. (2004). *A comparative study of housing reconstruction after two major earthquakes: The 1994 Northridge earthquake in the United States and the 1999 Chi-Chi earthquake in Taiwan*. PhD thesis, Texas A&M University.