SUPPLEMENTAL MATERIALS

ASCE *Journal of Water Resources Planning and Management*

Improving the Interpretation of Data-Driven Water Consumption Models via the Use of Social Norms

Renee Obringer, Roshanak Nateghi, Zhao Ma,
and Rohini Kumar

DOI: 10.1061/(ASCE)WR.1943-5452.0001611

Methods

The primary analysis was based on supervised learning (i.e., predictive modeling), a subset of statistical learning theory. The goal of supervised learning algorithms is to predict a target variable of interest based on a series of relevant predictor variables (i.e., the independent variables) (Hastie et al., 2009). Supervised learning algorithms often take one of two forms: parametric and non-parametric. The parametric models are based on some previously determined distributional assumption (e.g., multiple linear regression, ridge regression, etc.). While offering great interpretability, parametric models tend to be rigid in structure, with comparatively lower predictive power (Hastie et al., 2009). On the other hand, non-parametric algorithms make no assumptions about the data distribution and dependency structures and, therefore, tend to offer higher accuracy, though often at the cost of interpretability (Hastie et al., 2009). Certain non-parametric algorithms, such as artificial neural networks or support vector machines, can operate as 'black boxes'. To strike a balance between predictive power and interpretability, this study leverages an ensemble-of-trees method known as the random forest algorithm (Breiman, 2001).

Random Forest

Random forest is a tree-based ensemble method that builds B bootstrapped, de-correlated regression trees and then aggregates those trees to a single model (Breiman, 2001). The additional layers of randomness introduced in the random forest algorithm that leads to reduced correlation among the trees leads to further variance reduction and as a result improved performance over bagged-trees. The final model can be represented by the average of all the trees:

$$\hat{f}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{S1}$$

where T_b is the regression tree and B is the number of bootstrapping iterations.

Explanatory Sequential Mixed-Methods Study Design

A mixed-methods study design is an approach to research commonly found in the social sciences. The aim of the study is to integrate quantitative and qualitative data related to the study question (Creswell and Clark, 2011, 2017; Johnson et al., 2007). The actual design and implementation of mixed-methods research is fairly flexible, however. Researchers can, for example, conduct quantitative and qualitative studies simultaneously, then combine the results into a single interpretation, which is often referred to as concurrent triangulation design (Creswell and Clark,

2017). Alternatively, mixed-methods research can be sequential, meaning that the quantitative and qualitative analyses are performed such that one analysis informs the other (Creswell and Clark, 2017). In this study, we followed a sequential *explanatory* design, since the quantitative analysis informed the qualitative data collection and analysis, which was used to make the final interpretation of the quantitative analysis (Creswell and Clark, 2017). This is a particularly common approach to mixed-methods research on socio-environmental systems, in which researchers are finding that the rich text of qualitative data can be used to explain real-world deviations from idealized models (Elsawah et al., 2020).

Semi-Structured Interview Questions

Section 1: Awareness of water and/or electricity conservation programs

1. Have you heard of any programs offered by the utility company, city, state or other entity that encourage people to reduce their water and/or electricity use? Could you please describe these programs for me?
2. Have you heard of any initiatives specific to your neighborhood that involve water or electricity conservation?

Section 2: Personal beliefs regarding water and/or electricity conservation

1. Could you tell me about how you use water and electricity in and around your home, that is inside your home as well as any landscaping or outdoor activities that require water or electricity?
2. Could you describe the general bill-paying process in your place of residence?
3. Could you tell me about how you think about water and electricity conservation?
4. Can you think of a situation that would lead you to reduce your water and/or electricity use?

Section 3: Perceptions of others' beliefs regarding water and/or electricity conservation

1. Do you think your friends and neighbors think about water and electricity conservation in a similar way that you do?
2. Do you think your friends and neighbors in your area are doing anything related to water or electricity conservation?

3. Do you expect others, that is your friends, neighbors, or people in your neighborhood, to conserve water or electricity?

4. Do you feel others, that is your friends, neighbors, or people in your neighborhood, expect you to conserve water or electricity?

5. How would you react or feel if you found out that others, that is your friends, neighbors, or people in your neighborhood, were actively conserving water?

Table S1: Demographic variables from the 2018 American Community Survey considered in this study.

Variable Category	Variables Included in Category
Birth Rate	total birth rate; birth rate for people aged 15-19; birthrate for people aged 20-34, birthrate for people aged 35-50
Education Level	percent of the population that have: less than a high school education; a high school education; some college education; associates degree; bachelor's degree; post-bachelor's degree
Income Level	percent of the population with a household income of: less than $20,000; between $20,000 and $35,000; between $35,000 and $50,000; between $50,000 and $75,000; between $75,000 and $100,000; between $100,000 and $150,000; between $150,000 and $200,000; over $200,000
Household Unit Type	percent of population that is made up of: families; married couples w/o kids; single-parent families
House Type	percent of population that resides in: detached house; attached house; mobile home; miscellaneous dwelling
House Value	percent of population that resides in houses valued: less than $50,000; between $50,000 and $100,000; between $100,000 and $250,000; between $250,000 and $500,000; between $500,000 and $1,000,000; over $1,000,000
Language	percent of population whose primary language at home is: english; spanish; other European language; Asian language; other language
Marital Status	percent of population that identifies as: single; married; separated; widowed; divorced; other status
Place of Birth	percent of population that was born in: Europe; Asia; Africa; Oceania; Caribbean; Central America; South America; non-US North America
Age	percent of population that are: under 18 years old; 20-29; 30-29; 40-49; 50-64; over 65 years old
Race	percent of population that identify as: white; Black; Indigenous; Asian; Pacific Islander; Latino; other racial identity
Poverty Rate	Poverty rate
Work Commute	percent of population that travels to work via: single car; carpool; public transit; bicycle; walking; other transport; none (i.e., work from home)

Source: Data from US Census Bureau (2018)

Figures

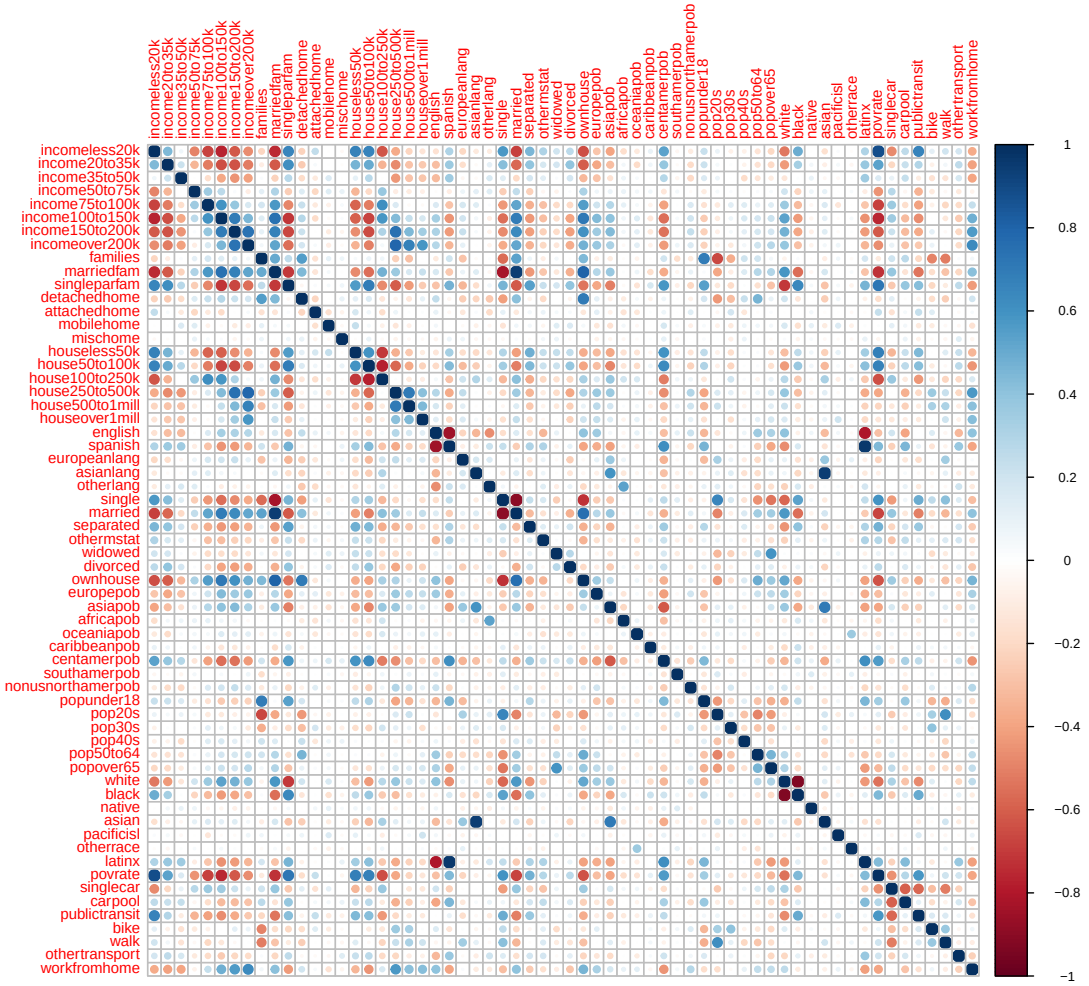

Figure S1: Correlation plot of the demographic variables.

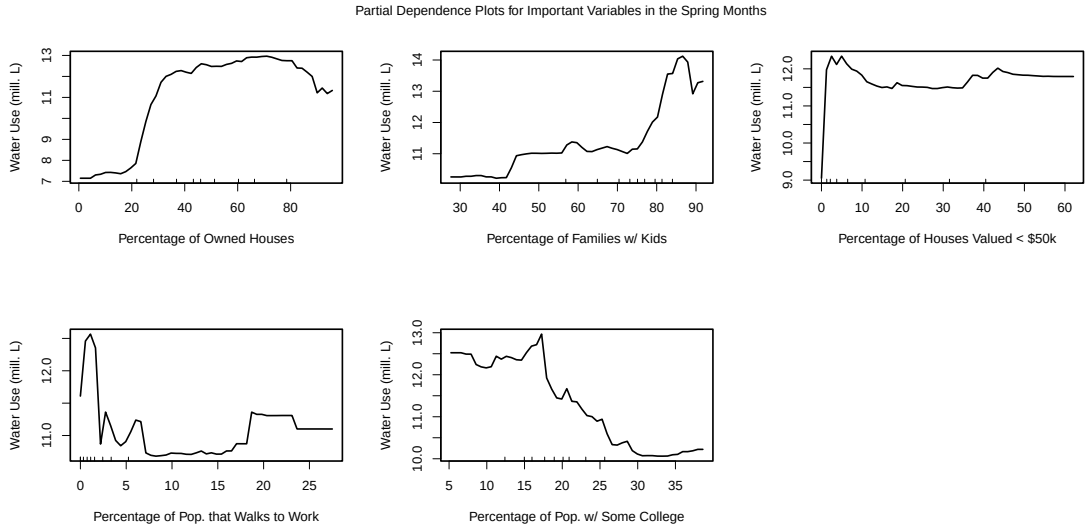

Figure S2: Partial dependence of the important variables in the spring months (moderate intensity analysis).

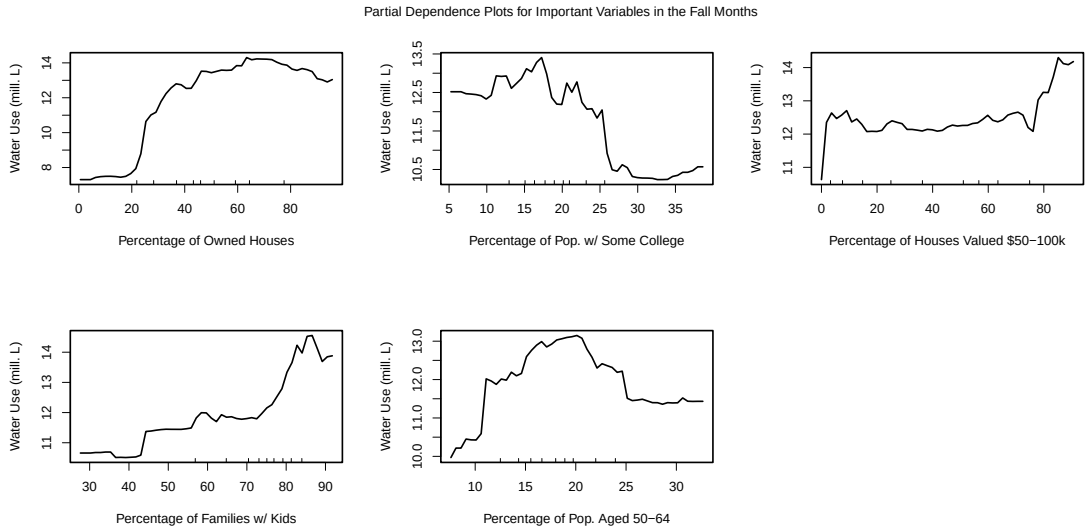

Figure S3: Partial dependence of the important variables in the fall months (moderate intensity analysis).

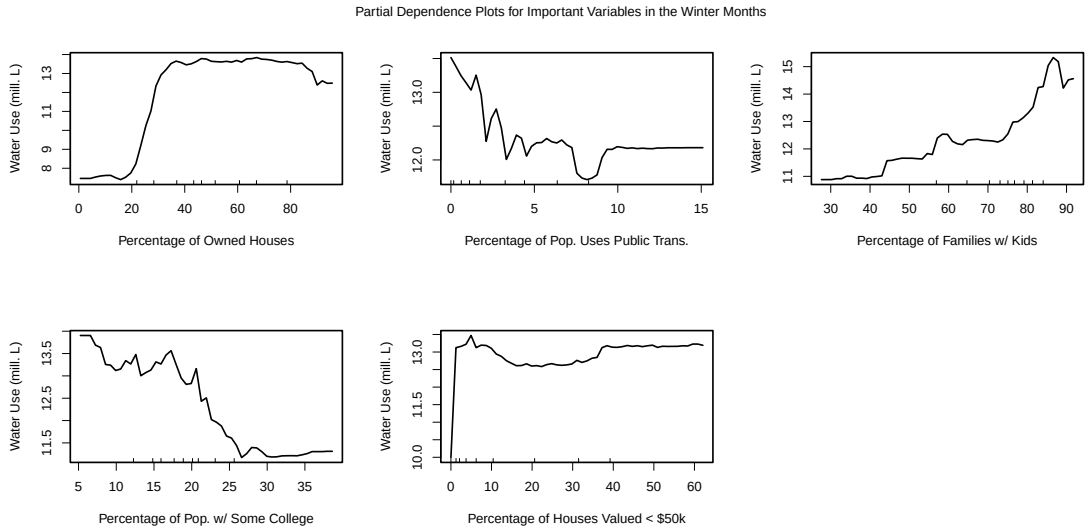

Figure S4: Partial dependence of the important variables in the winter months (moderate intensity analysis).

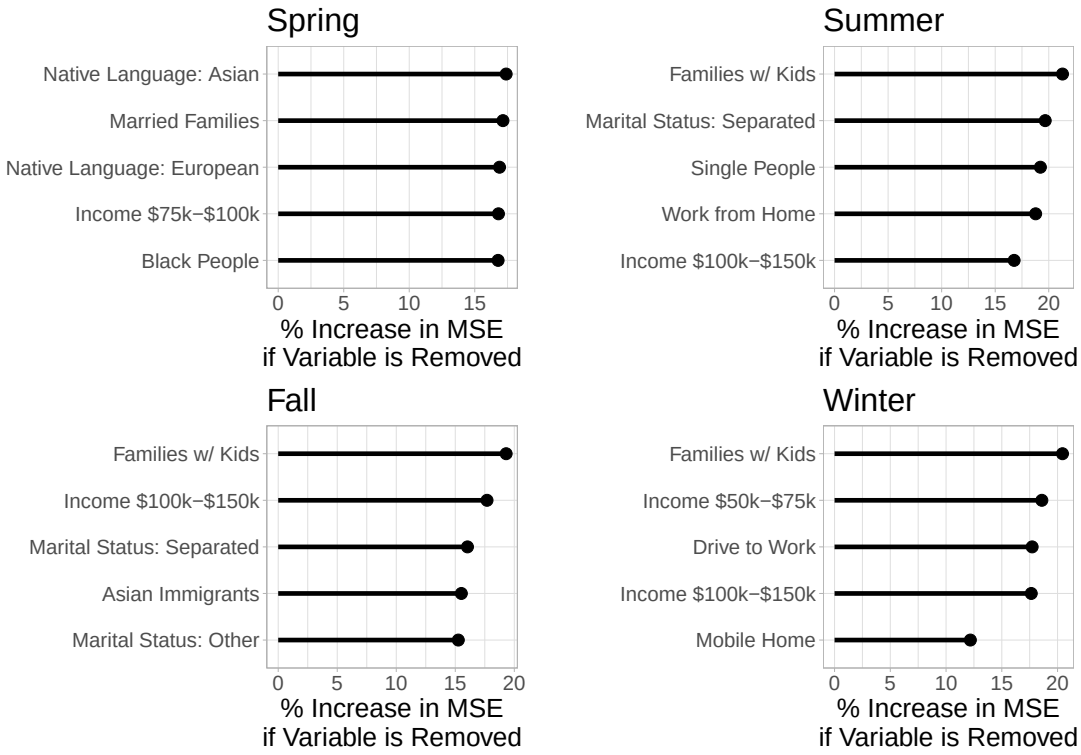

Figure S5: Important variables in the analysis of high intensity census tracts.

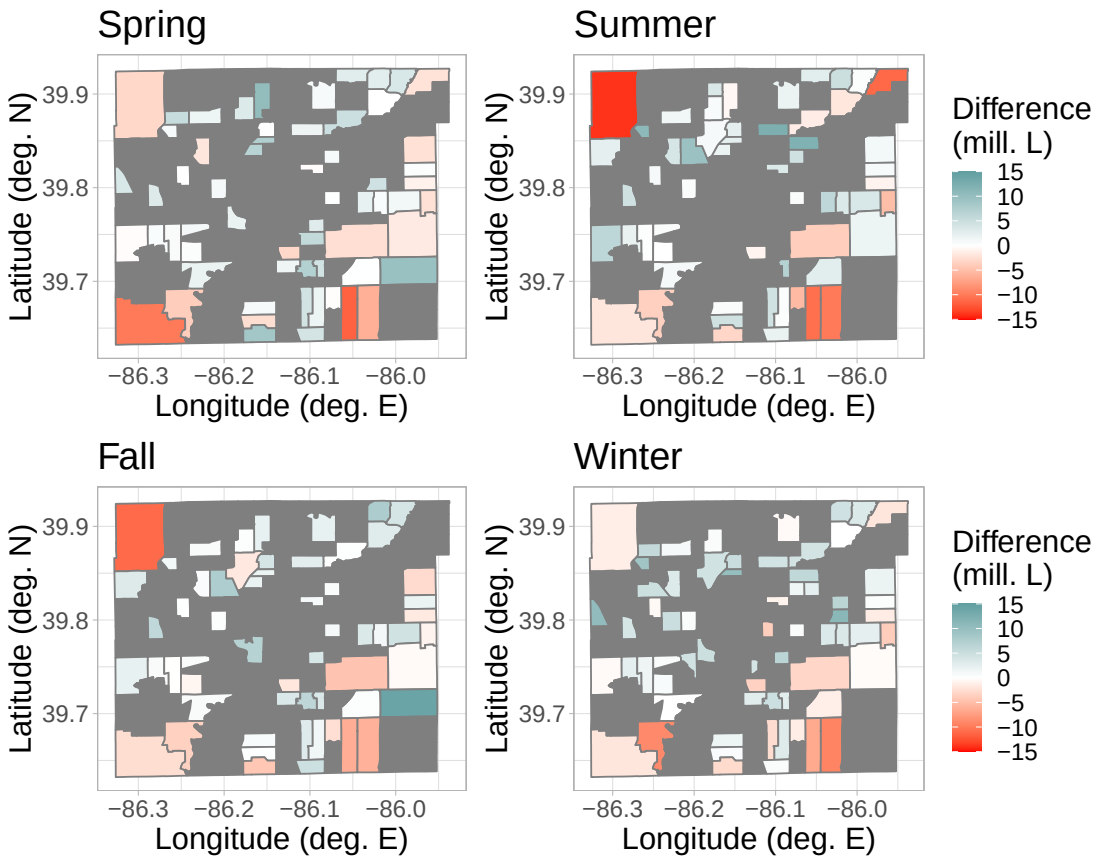

Figure S6: Differences between the actual and predicted values in the analysis of the high intensity census tracts.

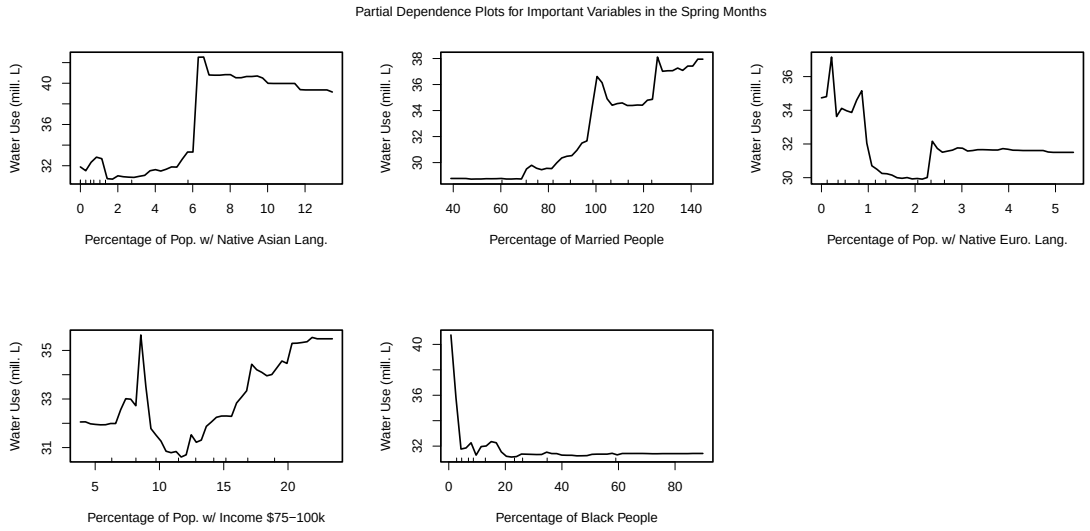

Figure S7: Partial dependence of the important variables in the spring months in the analysis of the high intensity census tracts.

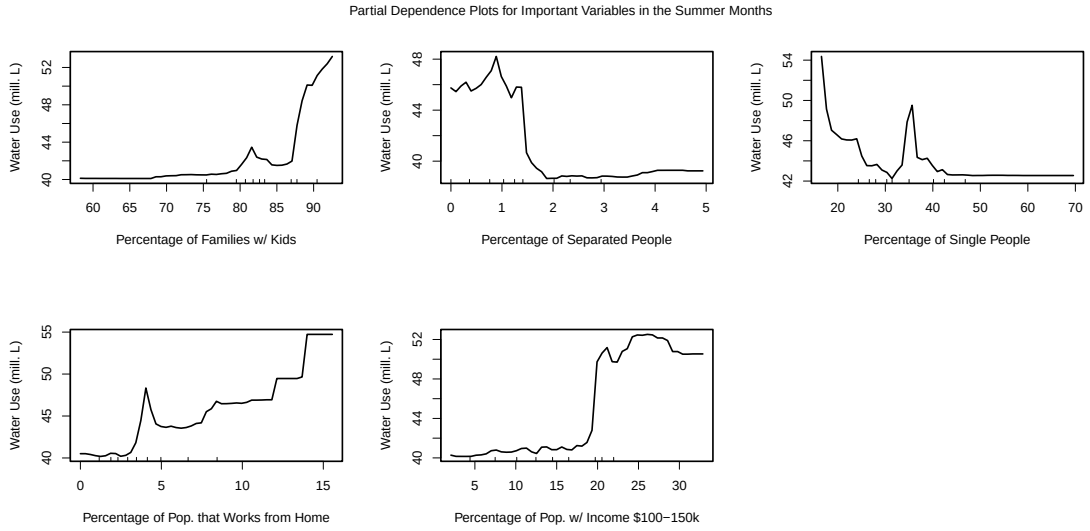

Figure S8: Partial dependence of the important variables in the summer months in the analysis of the high intensity census tracts.

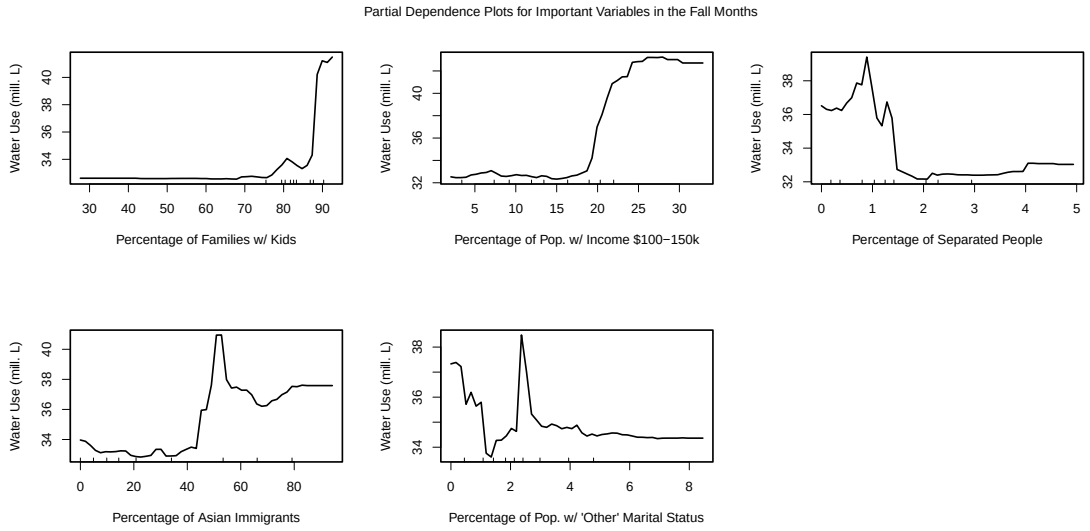

Figure S9: Partial dependence of the important variables in the fall months in the analysis of the high intensity census tracts.

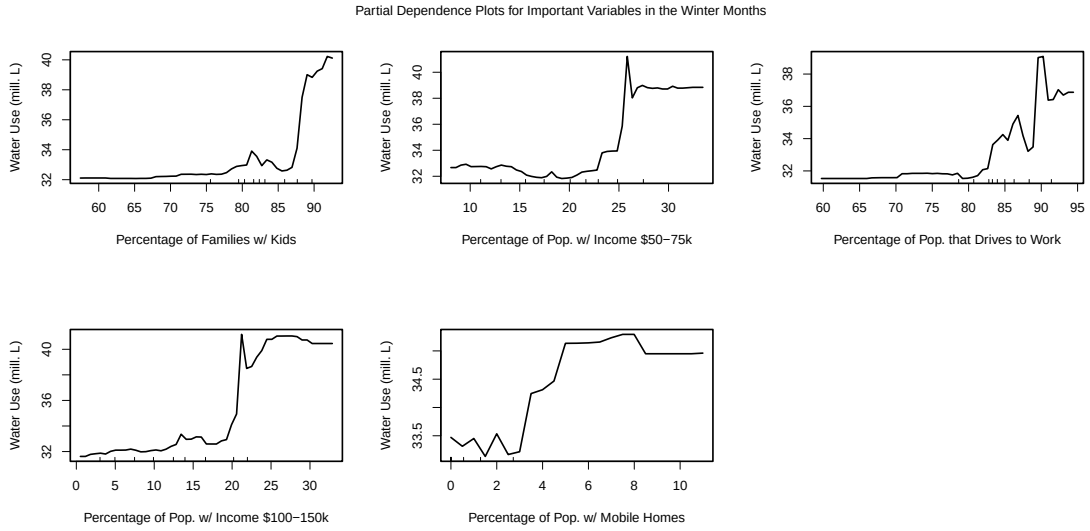

Figure S10: Partial dependence of the important variables in the winter months in the analysis of high intensity census tracts.

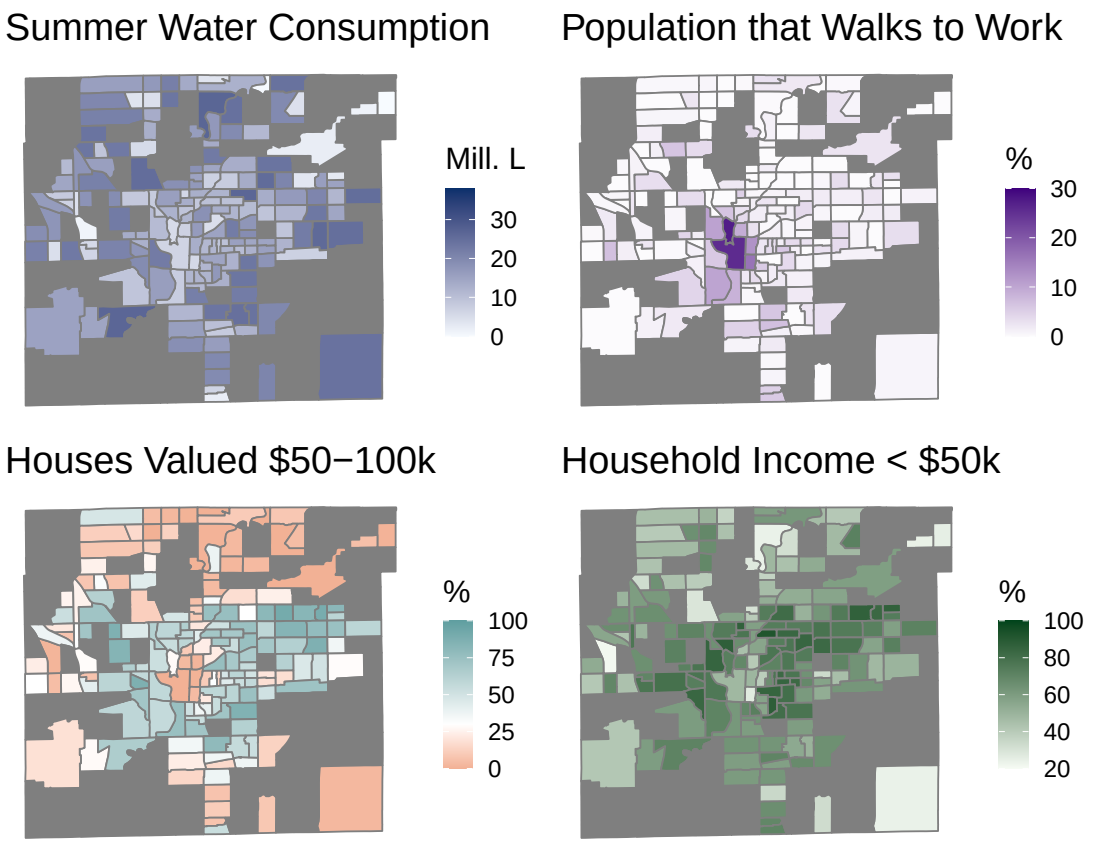

Figure S11: (a) Summer water consumption (million Liters). (b) The percentage of the population that walks to work. (c) The percentage of houses that are valued between $50,000 and $100,000, with an inflection point to show the 30% threshold. (d) The percentage of households with an income less than $50,000 a year.

References

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

John W. Creswell and Vicki L. Plano Clark. Choosing a Mixed Methods Design. In *Designing and Conducting Mixed Methods Research*. SAGE, 2011. ISBN 978-1-4129-7517-9.

John W. Creswell and Vicki L. Plano Clark. *Designing and Conducting Mixed Methods Research*. SAGE Publications, Thousand Oaks, CA, third edition, 2017.

Sondoss Elsawah, Tatiana Filatova, Anthony J. Jakeman, Albert J. Kettner, Moira L. Zellner, Ioannis N. Athanasiadis, Serena H. Hamilton, Robert L. Axtell, Daniel G. Brown, Jonathan M. Gilligan, Marco A. Janssen, Derek T. Robinson, Julie Rozenberg, Isaac I. T. Ullah, and Steve J. Lade. Eight grand challenges in socio-environmental systems modeling. *Socio-Environmental Systems Modelling*, 2:16226–16226, January 2020. ISSN 2663-3027. doi: 10.18174/sesmo.2020a16226.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media, August 2009. ISBN 978-0-387-84858-7.

R. Burke Johnson, Anthony J. Onwuegbuzie, and Lisa A. Turner. Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, 1(2): 112–133, April 2007. ISSN 1558-6898. doi: 10.1177/1558689806298224.