Using Random Forests and Geographic Weighted Regression to Assess Influential Variables on the Annual Energy Use Intensity of Residential Buildings in Portland, Oregon

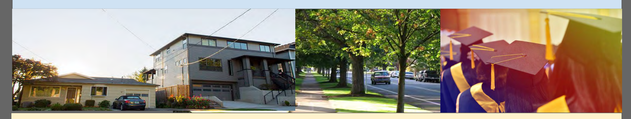
Zachary Neumann, Kristen Purdy, Alec Trusty

Portland State UNIVERSITY

EnergyTrust of Oregon

Introduction:

- This was an exploratory analysis project to find which metrics influence annual residential energy intensity in Portland, Oregon
- Understanding which factors influence energy consumption can help state and local governments achieve goals to reduce CO2 emissions
- We examined socio-economic, structural, and environmental factors that were significant predictors in prior research for modelling energy consumption (Escobedo, Seitz, and Zipperer, 2012; Ewing et al., 2008; Huebner et al., 2015)
- We used Random Forests to find influential variables, and geographic weighted regression to model predicted energy intensity based on the influential variables from Random Forests



Data Processing & Integration:

Primary Data Sources:

Socio-economic factors:

- 2014 Census Data via Social Explorer, US Census Bureau

Building-Level factors:

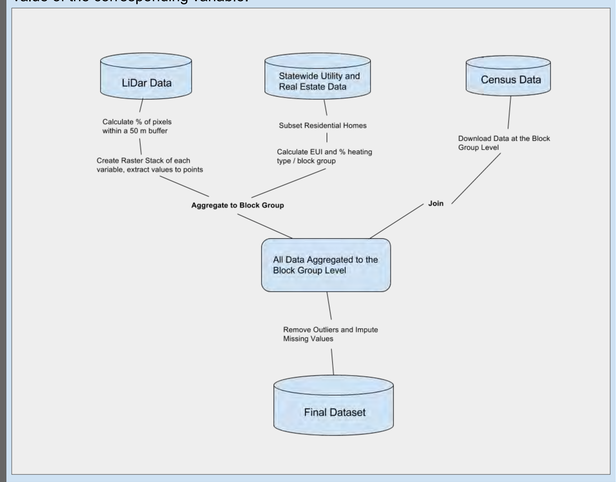
- Real Estate and Energy Consumption Data (2012) from the Sustaining Urban Places Research Lab's (SUPR) Integrated Energy Consumption Dataset

Environmental factors:

- 1m Lidar-derived canopy cover, vegetation, and biomass index from the SUPR Lab (2014)
- Digital Elevation Model (DEM) from the Regional Land Inventory System (RLIS) were used to derive slope and aspect

Data Integration:

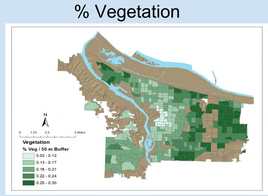
We used R and ArcGIS for all data processing, integration, and analysis. Residential homes were subset from the energy consumption dataset. Energy Use Intensity (EUI)–total annual kWH/sq foot—of each home was calculated, along with the percentage of homes within each block group that use electric, oil, or gas heating sources. The percentage of canopy cover, vegetation, as well as biomass index were calculated within a 50 meter buffer of the building address centroid. All building-level data was aggregated to the block group, and the census data was joined to the aggregated dataset. Outliers were removed and missing values were imputed using the median value of the corresponding variable.



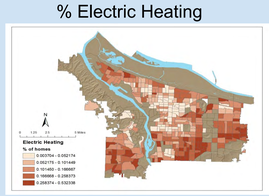
Results:

Six Explanatory Variables Found in Random Forest used for GWR

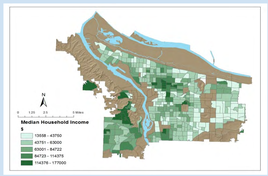
% Vegetation



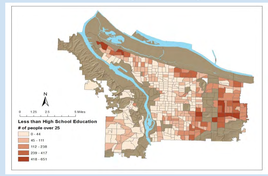
% Electric Heating



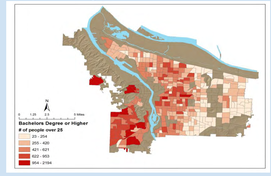
Median Household Income



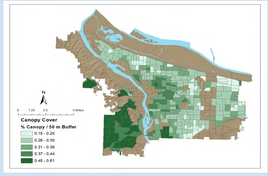
Bachelors Degree or Higher



High School or Lower

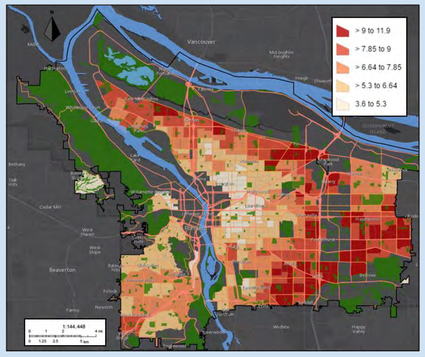


% Canopy Cover

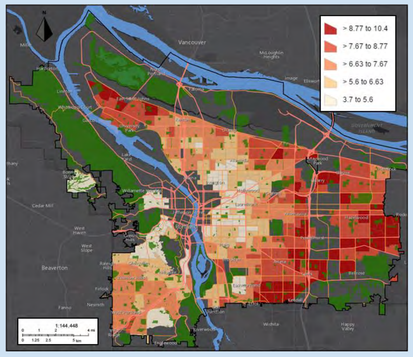


Geographically Weighted Regression

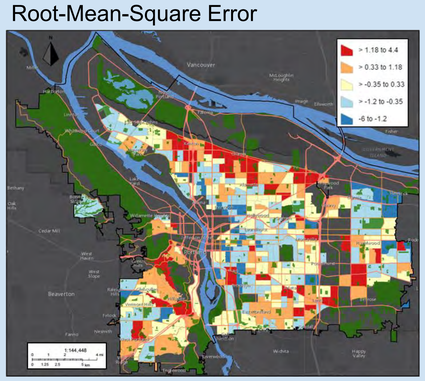
Observed Energy Use Intensity



Predicted Energy Use Intensity



Root-Mean-Square Error



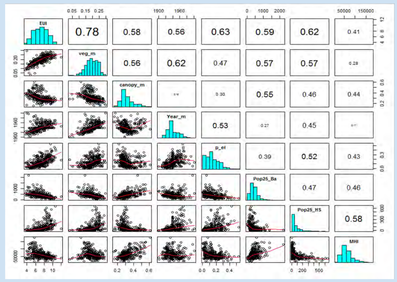
Above are maps of observed annual residential energy use intensity (kWH/sq. foot) and predicted annual residential energy use intensity for comparison. The map of root-mean-square error (left) shows the amount that the observed and predicted values differed when using the six explanatory variables (above) in a GWR (R2 = 0.716). Generally, we see the model overpredicted on the outer NE and SE edges of Portland, and under predicted in inner E Portland. This could represent a key variable missing in these regions, perhaps related to zoning.

Analysis methods:

20 Explanatory Variables Used in Random Forest

Category	Alias	Variable
Building Structure	Year_m	Average Year House was Built
	Sqft_m	Average size of the home
	P_gas	Percentage of homes heated by gas
	P_oil	Percentage of home heated by oil
	P_el	Percentage of homes with electric heating
Household	Hsize_m	Average household size
	MHI	Median household Income
Surrounding Landscape	Aspect	Dominant direction of slope
	Canopy_m	Average percentage canopy cover around each home
	Veg_m	Average percentage vegetation
	Bio_m	Average biomass Index
Neighborhood	OOHU	Owner occupied housing units
	CivPop	Civilian population
	CivPop_Emp	Civilian population, employed
	CivPop_Un	Civilian population, unemployed
	Pop25	population over 25 years old
	Pop25_Ba	population over 25, with education higher than bachelors
	T_Pop	total population
	Pop_Den_sq	population density per square mile
	T_Pop_17	total population under 17 years old
	T_Pop_5	total population 5 years old

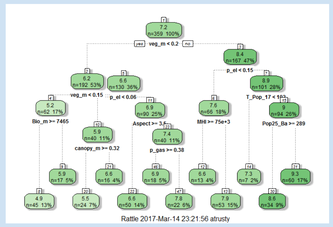
Correlation Matrix of EUI and Explanatory Variables



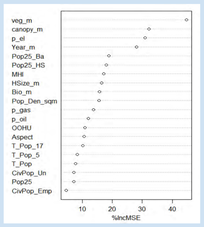
Regression Tree & Random Forest

A decision tree and machine learning technique, Random Forests (Brieman, 2016), was applied to determine the most influential explanatory variables. The model, which was optimized at 1000 trees and 5 randomly selected variables at each node, resulted in a mean square error of 0.58 and 77.23% variance explained.

Single Regression Tree



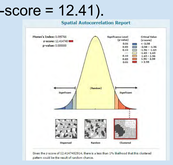
Variable Importance Plot



Geographically Weighted Regression (GWR)

We used a local statistical technique, GWR, to assess where our variables were predicting EUI the best, and where they weren't. Year_m was highly correlated with Veg_m, and removed to reduce multi-colinearity in the explanatory variables. Veg_m, Canopy_m, P_el, Pop25_Ba, Pop25_HS, and MHI were used. GWR was run with a fixed kernel type and AIC bandwidth method. Global Moran's I was used to check for spatial autocorrelation of the residuals, and was found to be spatially clustered (z-score = 12.41).

Model	Variance Explained
Random Forest	0.7723
Geographically weighted Regression	0.716



Discussion:

- Factors that decreased energy consumption were: canopy cover, educational attainment bachelor's degree or more, median household income
- Factors that increased energy consumption were: vegetation, educational attainment less than high school, electric heating
- Generally we see the model overpredicted in outer NE and SE Portland, and under predicted in inner E Portland
- We tried to see what other variables could be added to our model to improve our results
- One variable we examined was homes built after 1985. We found that this may help explain energy use in E Portland, but the pattern did not line up with the trend seen in NE Portland
- The model could be improved by looking at areas where the EUI was over or under predicted to find additional key variables, perhaps related to urban planning and zoning
- Once improved, these findings could be used for planning efforts to help decrease energy consumption

References:

Ewing, Reid, and Fang Rong. "The impact of urban form on US residential energy use." Housing policy debate 19.1 (2008): 1-30.

Escobedo, F., Seitz, J., & Zipperer, W. (2012). The Effect of Gainesville's Urban Trees on Energy Use of Residential Buildings. IFAS Extension. Retrieved February 2, 2017, from https://edis.ifas.ufl.edu/pdffiles/FR/FR27300.pd

Huebner,G. M., Hamilton, I., Chalabi, Z., Shipworth, D., & Oreszczyn, T. (2015). Explaining domestic energy consumption–the comparative contribution of building factors, socio-demographics, behaviours and attitudes. *Applied energy*, *159*, 589-600.

Dietz, R. (2015, May 19). New Single-Family Home Size Increases at the Start of 2015. Retrieved March 19, 2017, from http://eyeonhousing.org/2015/05/new-single-family-home-size-increases-at-the-start-of-2015/

Random Forests Leo Breiman and Adele Cutler. (n.d.). Retrieved March 19, 2017, from https://www.stat.berkeley.edu/~breiman/RandomForests/