

Summary of Prediction of Housing Satisfaction

Xunhang (David) Gao

August 2025

Contents

1	Introduction	2
2	Data	2
2.1	Variable Definitions	2
3	Housing Models	3
3.1	Notation	3
3.2	GAM models	3
3.3	Random Forest Models	4
4	Results	5
4.1	GAM models	5
4.2	Random Forest models	5
4.3	Prediction of Expected Survey Results	5
4.4	Prediction of Housing Satisfaction	11
4.5	Supplementary Models for Small Communities	14
4.6	Interpretation of Supplementary GAM model	16
5	References	22

[Click here to visit Latex version](#)

1 Introduction

Extending the work of Batista et al. (2023) [1], which predicts resident satisfaction with public services using publicly available data, this project focuses on modeling housing satisfaction in Iowa's small towns using survey data from the Iowa Small Towns Project (ISTP) and government sources.

2 Data

Objective: Estimate the average housing satisfaction across 1,008 cities in Iowa.

Data Source 1: Iowa Small Towns Project (ISTP) survey randomly sampled 99 cities from a total of 1,008 cities and collected responses from 100 individuals in each selected city.

Data Source 2: Publicly available data from Iowa.gov and the American Community Survey (ACS).

2.1 Variable Definitions

2.1.1 Survey Covariates

Age: variable range from 18 to 100;

Safe: variable ranges from 1 to 7, where 1 represents “safe”, and 7 indicates “dangerous”;

Place of Residence: variable range from 1 to 3, where 1 represents “within the city limits”, “2 presents “outside the city, on a farm”, and 3 represents “outside the city, not on a farm”;

Condition of Parks Importance: variable representing importance of condition of parks to respondent, ranging from 1 to 2, where 1 as “No”, and 2 as “Yes”;

Unemployed: variable ranges from 0 to 1, where 0 represents “employed”, and 1 represents “unemployed”;

Public Schools Importance: variable representing importance of the public schools to respondent, ranging from 1 to 2, where 1 as “No”, and 2 as “Yes”;

2.1.2 City-level Covariates

acs.higherEd_ma3: proportion of population 25 years and over having Bachelor’s degree or higher;

acs.unemployment_ma3: proportion of unemployed civilian labor force;

acs.workAdults_ma3: proportion of people 16 years old and over;

acs.ownerOcc_perHouse_ma3: number of owner occupied housing units;

housing.homevalueToIncome_neigh_ma3: median monthly housing costs as a percentage of median household income;

NatAmen.Natural_amenity_Scale: a measure of the physical characteristics of a county area that enhance the location as a place to live;

county.Limited access to healthy foods_% Limited Access: Percentage of population who are low-income and do not live close to a grocery store;

county.Inadequate social support_% No Social-Emotional Support: Percentage of adults without adequate social/emotional support;

Crime.crime rate idx log: log of total crime per capita;

3 Housing Models

3.1 Notation

Parameter of interest: $y_{i,j} \in \{1, 2, 3, 4\}$ represents the 4-level ordered housing satisfaction of individual j in city i .

Survey covariate: Relevant individual-level covariates were collected in the survey, denoted by $x_{i,j} = [x_{i,j,1}, \dots, x_{i,j,p}]^T \in \mathbb{R}^p$, for individual j in city i .

City covariate: Publicly available city-level data, such as the employment rate, denoted by $z_i = [z_{i,1}, \dots, z_{i,q}]^T \in \mathbb{R}^q$.

3.2 GAM models

We consider the following GAM model for housing satisfaction prediction for cities which survey is conducted,

$$y_{i,j} = x_{i,j}^T \beta + h(z_i, \eta_h) + \epsilon_{i,j}, \quad (1)$$

where $\mathbb{E}(\epsilon_{i,j}) = 0$.

And prediction of expected housing satisfaction for city i without available data from survey follows

$$\bar{y}_{i,.} = \bar{x}_{i,.}^T \beta + h(z_i, \eta_h) + \epsilon_{i,.}, \quad (2)$$

where $\bar{x}_{i,.}$ is the city-level average value of covariate $x_{i,j}$ across individuals.

We fit the models using the `mgcv:::gam` package in R (Wood, S. N. (2020) [2]).

3.2.1 Two stage Model

The $\bar{x}_{i,.}$ average value of survey covariate $x_{i,j}$ across individuals, is not available if survey is not conducted in city i . And the solution is predicting the expected value of survey covariates using city-level data, $z_i \in \mathbb{R}^q$.

We consider the following GAM model for prediction of expected values of survey covariates,

$$\bar{x}_{i,.} = g(z_i, \eta_g) + u_i, \quad (3)$$

where $\mathbb{E}(u_i) = 0$.

3.2.2 Model Selection

The variable selection of model 1 and model 2 is conducted using stepwise selection based on MSE computed from LOOCV. And in each step of variable selection, the model selection is processed through maximizing that penalized likelihood with `mgcv:::gam` in R.

In the second stage model 3 of predicting selected survey covariates, the pre-screening of city-level covariates z_i is processed with LASSO first. Further variable selection is performed using `regsubsets` function from `leaps` package, returning the best models at each subset size $k \in \{1, 2, \dots, nvmax\}$. Then, the model selection is carried out based on the MSE calculated from LOOCV with `mgcv:::gam`.

3.3 Random Forest Models

Similar to GAM models, we predict housing satisfaction by chaining two random-forest stages, one classification random forest model for prediction of housing satisfaction and another regression random forest model for prediction of selected survey covariates.

3.3.1 Stage 1: individual-level classification random forest

Train a random forest on $\{(\mathbf{x}_{i,j}, y_{i,j})\}$ over surveyed cities. For any selected survey covariates \mathbf{x} , the forest provides out-of-bag (oob) class probabilities

$$\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), p_3(\mathbf{x}), p_4(\mathbf{x})), \quad \sum_{k=1}^4 p_k(\mathbf{x}) = 1. \quad (4)$$

Define the expected (ordinal) score

$$s(\mathbf{x}) = \sum_{k=1}^4 k p_k(\mathbf{x}). \quad (5)$$

For a surveyed city i , aggregate individual scores:

$$\hat{y}_{i,\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} s(\mathbf{x}_{i,j}). \quad (6)$$

We apply permutation variable importance (vimp) computed from oob predictions and perform backward elimination, selecting the smallest subset whose mean oob error is within one standard error of the minimum (the 1-se rule), using VSURF (Genuer, R. and Poggi, J.M. and Tuleau-Malot, C. (2015) [3]) package and varSelRF (Diaz-Uriarte, R. and Alvarez de Andres, S. (2005) [4]) package in R.

3.3.2 Stage 2: city-level regression random forests

For each retained survey covariate index r , compute surveyed-city means

$$\bar{x}_{i,r} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j,r}. \quad (7)$$

Fit a regression random forest

$$g_r : \mathbf{z}_i \mapsto \bar{x}_{i,r}, \quad (8)$$

assessed by oob mean squared error. Apply the same oob–vimp elimination with the 1–se rule on the city-level covariates \mathbf{z} as in stage 1.

For a non-surveyed city i , form the expected survey vector

$$\hat{\mathbf{x}}_{i,\cdot} = (g_r(\mathbf{z}_i))_{r \in \mathcal{S}}, \quad (9)$$

where \mathcal{S} indexes the survey covariates retained in stage 1.

3.3.3 Final Prediction

Evaluate the stage 1 classifier at $\hat{\mathbf{x}}_{i,\cdot}$:

$$\mathbf{p}(\hat{\mathbf{x}}_{i,\cdot}) = (p_k(\hat{\mathbf{x}}_{i,\cdot}))_{k=1}^4, \quad \hat{y}_{i,\cdot}^* = \sum_{k=1}^4 k p_k(\hat{\mathbf{x}}_{i,\cdot}).$$

Report $\hat{y}_{i,\cdot}$ for surveyed cities and \hat{s}_i^* otherwise.

We fit the models using the `ranger` package in R (Wright, M. N. & Ziegler, A. (2017) [5]).

4 Results

4.1 GAM models

The final GAM model of predicting expected housing satisfaction follows

$$y_{i,j} = \beta_0 + \beta_s x_{i,j,s} + \beta_{pr} x_{i,j,pr} + \beta_{cpi} x_{i,j,cpi} + \beta_u x_{i,j,u} + \beta_{psi} x_{i,j,psi} + \sum_{k=1}^9 \eta_k h_k(z_{i,k}) \quad (10)$$

for individual j in city i with survey data, and

$$\bar{y}_{i,\cdot} = \beta_0 + \beta_s \bar{x}_{i,\cdot,s} + \beta_{pr} \bar{x}_{i,\cdot,pr} + \beta_{cpi} \bar{x}_{i,\cdot,cpi} + \beta_u \bar{x}_{i,\cdot,u} + \beta_{psi} \bar{x}_{i,\cdot,psi} + \sum_{k=1}^9 \eta_k h_k(z_{i,k}) \quad (11)$$

for city i without survey data, where

$x_{i,j,s}$: safe;

$x_{i,j,pr}$: place of residence;

$x_{i,j,cpi}$: condition of parks importance;

$x_{i,j,u}$: unemployed;

$x_{i,j,psi}$: public schools importance;

$z_{i,k}$ for $i \in \{1, \dots, 5\}$ are city-level covariates, $\{\text{acs.higherEd_ma3}, \text{acs.unemployment_ma3}, \text{acs.workAdults_ma3}, \text{acs.ownerOcc_perHouse_ma3}, \text{housing.homevalueToIncome_neigh_ma3}\}$.

4.2 Random Forest models

The final random forest model 4 selected for prediction of expected housing satisfaction of city i include survey covariates,

$x_{i,j,s}$: safe;

$x_{i,j,pr}$: place of residence;

$x_{i,j,age}$: Age;

$z_{i,k}$ for $i \in \{1, \dots, 5\}$ are city-level covariates, $\{\text{acs.higherEd_ma3}, \text{acs.unemployment_ma3}, \text{acs.workAdults_ma3}, \text{acs.ownerOcc_perHouse_ma3}, \text{housing.homevalueToIncome_neigh_ma3}\}$.

4.3 Prediction of Expected Survey Results

4.3.1 Safe (GAM and Random Forest models)

The GAM model of prediction of expected value of safe for city i follows

$$\bar{x}_{i,\cdot,s} = \sum_{k=1}^8 \eta_k g_k(z_{i,k}), \quad (12)$$

where $z_{i,k}$ for $k \in \{1, \dots, 8\}$ are city-level covariates `higherEd_ma3`, `pop25_ma3`, `workAdults`, `ownerOcc_perHouse`, `enrollment_ELL`, `enrollment_black_ma3`, `taxtVsISASP_ma2`, `netOpenEnroll_norm`.

Figure 1 presents the estimated final model 12 of predicting the expected values of `Safe` in city i .

The random forest model of prediction of expected value of safe for city i include city level covariates $z_{i,k}$ for $k \in \{1, \dots, 5\}$, higherEd_ma3, freeLunch, poverty, establishments_perCap_ma2, exp_perPupil_tot_real.

Figure 2 presents the comparison between distributions of observed and predicted expected values of Safe within the 99 surveyed cities for GAM model and random forest model. GAM model reaches the MSE of 0.04790531 and the random forest model reaches the MSE of 0.09524229.

Figure 3 presents the prediction of Safe across all cities in Iowa State by both GAM model and random forest model.

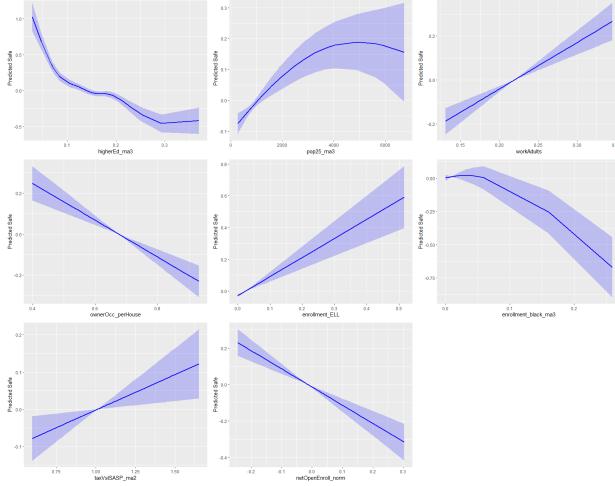


Figure 1: The GAM model of predicting Safe.

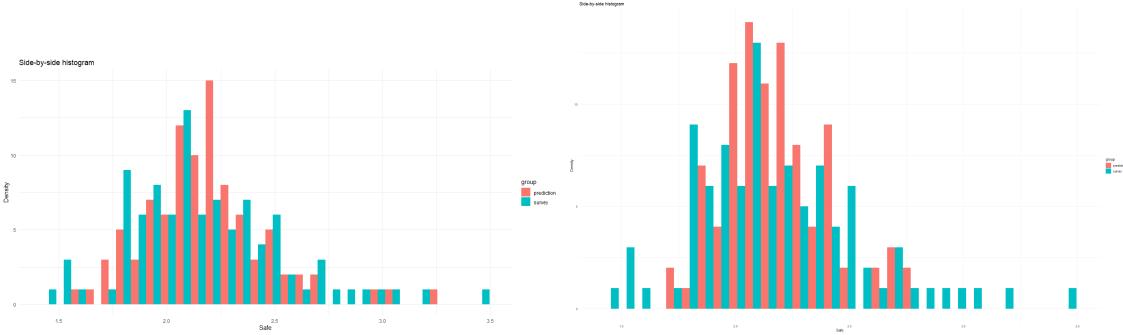


Figure 2: The distribution of predicted expected Safe from GAM model (left) and random forest model (right) approximately follows the observed distribution. But GAM model performs better on prediction of extreme cases.

4.3.2 Place of Residence (GAM and Random Forest models)

The GAM model of predicting value of Place of Residence for city i follows

$$\bar{x}_{i,\cdot,pr} = \sum_{k=1}^6 \eta_k g_k(z_{i,k}), \quad (13)$$

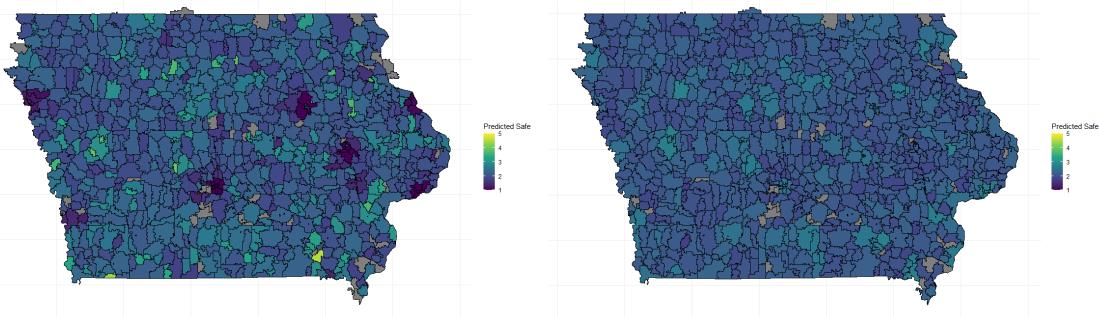


Figure 3: The distribution of predicted expected Safe of both GAM model (left) and random forest model (right) follows similar patterns, but predictions based on random forest model is more concentrated.

where $z_{i,k}$ for $k \in \{1, \dots, 6\}$ are city-level covariates `povety`, `rentToIncome_ma3`, `costToIncome`, `newAdults_perCap_cagr2`, `employees_reconcile_perCap`, `enrollment`.

Figure 4 presents the estimated GAM final model 13 of predicting the expected values of Place of Residence in city i .

The random forest model of prediction of expected value of safe for city i include city level covariates $z_{i,k}$ for $k \in \{1, \dots, 4\}$, `employees_reconcile_perCap`, `povety`, `establishments_neigh_perCap`, `workAdults_cagr2`.

Figure 5 presents the comparison between distributions of observed and predicted expected values of Place of Residence within the 99 surveyed cities for GAM model and random forest model. GAM model reaches the MSE of 0.0224982 and the random forest model reaches the MSE of 0.03937562.

Figure 6 presents the prediction of Place of Residence across all cities in Iowa State by both GAM model and random forest model.

4.3.3 Age (Random Forest models)

The random forest model of prediction of expected value of safe for city i include city level covariates $z_{i,k}$ for $k \in \{1, \dots, 4\}$, `establishments_perCap_ma3`, `freeLunch`, `housingCost_real`, `taxVsISASP`.

Figure 7 presents the comparison of distributions of observed and predicted expected Age values within the 99 surveyed cities with MSE as 5.784371 and the prediction of Age across all cities in Iowa State.

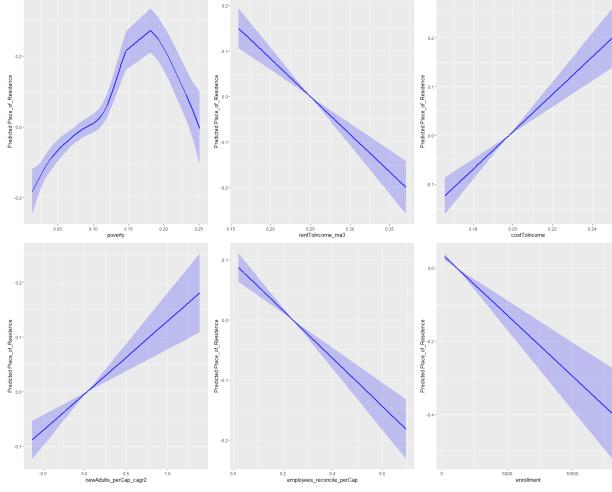


Figure 4: The GAM model of predicting Place of Residence.

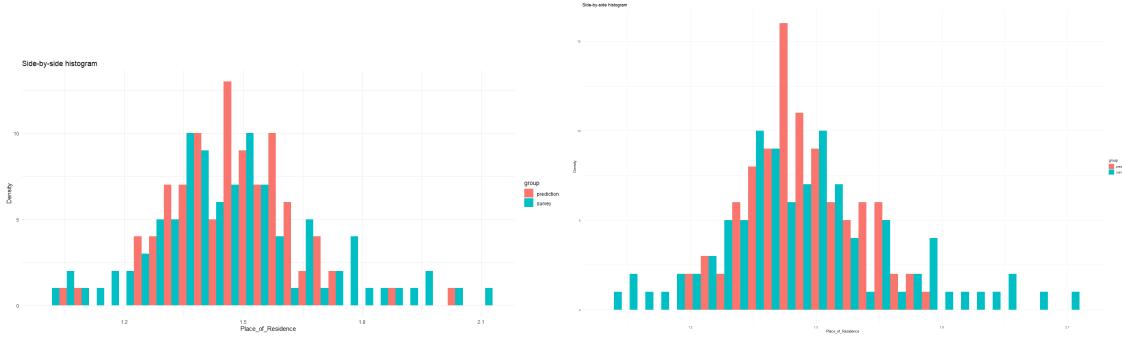


Figure 5: The distribution of predicted expected Place of Residence approximately follows the observed distribution.

4.3.4 Condition of Parks Importance (GAM model)

The model of predicting value of Condition of Parks Importance for city i follows

$$\bar{x}_{i,\cdot,cpi} = \sum_{k=1}^6 \eta_k g_k(z_{i,k}), \quad (14)$$

where $z_{i,k}$ for $k \in \{1, \dots, 6\}$ are city-level covariates `povety`, `rentToIncome_ma3`, `costToIncome`, `newAdults_perCap_cagr2`, `employees_reconcile_perCap`, `enrollment`.

Figure 8 presents the estimated final model 14 of predicting the expected values of Condition of Parks Importance in city i .

Figure 9 presents the comparison of distributions of observed and predicted expected Condition of Parks Importance values within the 99 surveyed cities with MSE as 0.001420365 and the prediction of Condition of Parks Importance across all cities in Iowa State.

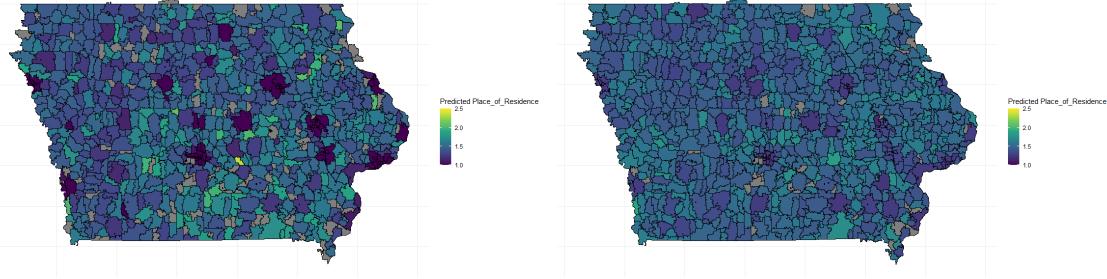


Figure 6: The distributions of predicted expected Place of Residence of both GAM model (left) and random forest model (right) follows similar patterns, but predictions based on random forest model is more concentrated.

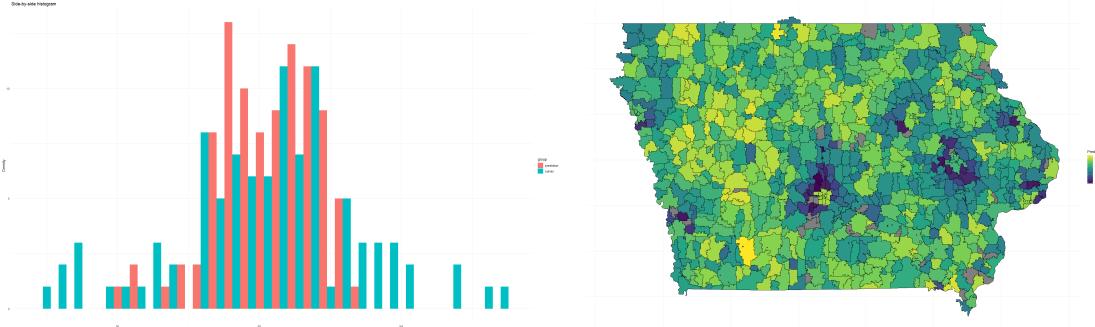


Figure 7: The distribution of predicted expected Condition of Parks Importance approximately follows the observed distribution.

4.3.5 Unemployed (GAM model)

The model of predicting value of Unemployed for city i follows

$$\bar{x}_{i,\cdot,u} = \sum_{k=1}^6 \eta_k g_k(z_{i,k}), \quad (15)$$

where $z_{i,k}$ for $k \in \{1, \dots, 6\}$ are city-level covariates `pop25_cagr3`, `costToIncome_ma3`, `Capital_Projects_perCap_real`, `Capital_Projects_perCap_real_log`, `employees_reconcile_neigh_perCap_weighted_avg_ma3_cagr2`, `closure_impact`.

Figure 10 presents the estimated final model 15 of predicting the expected values of Unemployed in city i .

Figure 11 presents the comparison of distributions of observed and predicted expected Unemployed values within the 99 surveyed cities with MSE as 0.0002879776 and the prediction of Unemployed across all cities in Iowa

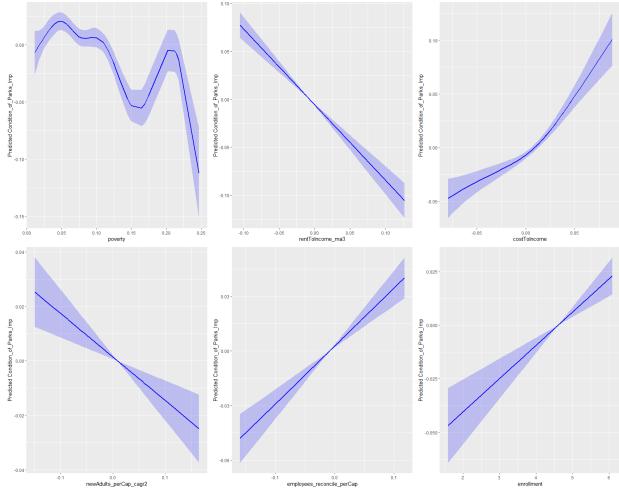


Figure 8: The GAM model of predicting Condition of Parks Importance.

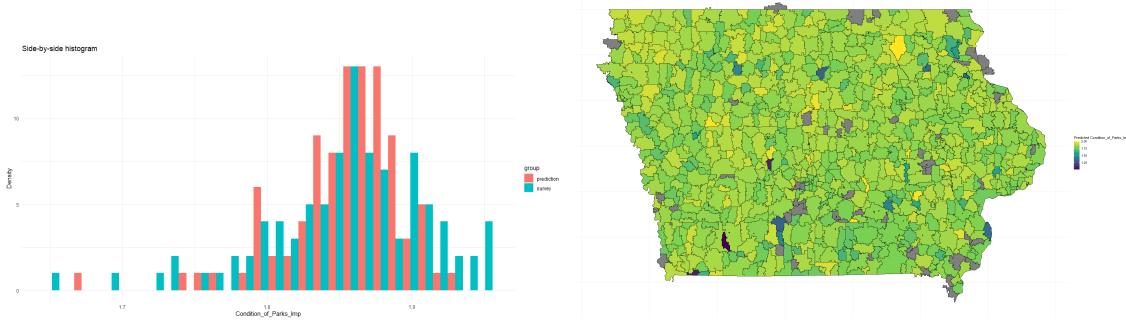


Figure 9: The distribution of predicted expected Condition of Parks Importance approximately follows the observed distribution.

State.

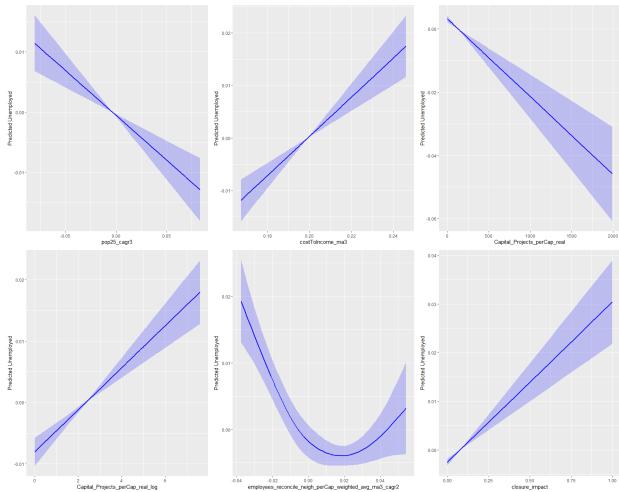


Figure 10: The GAM model of predicting Unemployed.

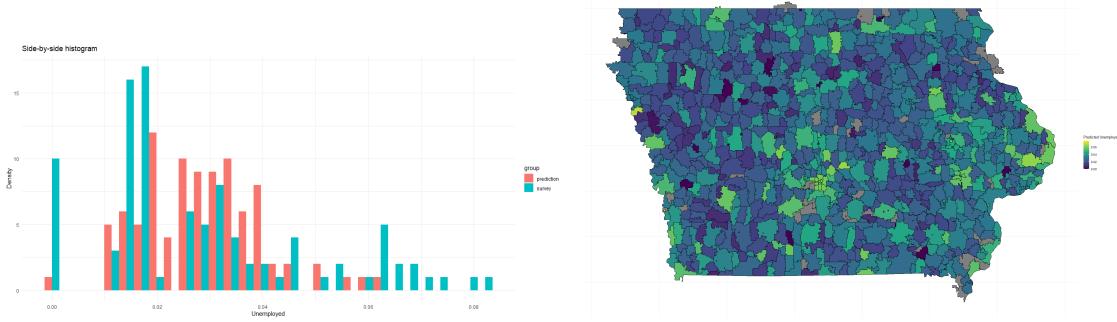


Figure 11: The distribution of predicted expected Unemployed approximately follows the observed distribution.

4.3.6 Public Schools Importance (GAM model)

The model of predicting value of Public Schools Importance for city i follows

$$\bar{x}_{i,\cdot,psi} = \sum_{k=1}^2 \eta_k g_k(z_{i,k}), \quad (16)$$

where $z_{i,k}$ for $k \in \{1, 2\}$ are city-level covariates `enrollment_ma3`, `ISASP_ma3`.

Figure 12 presents the estimated final model 16 of predicting the expected values of Public Schools Importance in city i .

Figure 13 presents the comparison of distributions of observed and predicted expected Public Schools Importance values within the 99 surveyed cities with MSE as 0.004145587 and the prediction of Public Schools Importance across all cities in Iowa State.

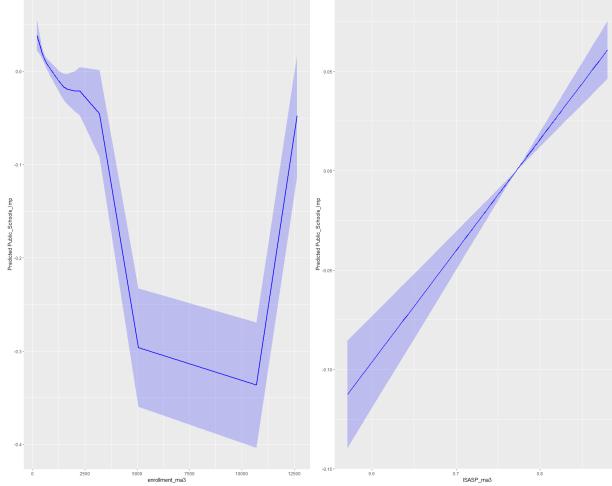


Figure 12: The GAM model of predicting Public Schools Importance.

4.4 Prediction of Housing Satisfaction

Figure 14 presents the prediction of housing satisfaction of the 99 surveyed cities based on ordinary linear model with selected survey covariates in final GAM model And the table 1 presents the estimation of the ordinary linear model,

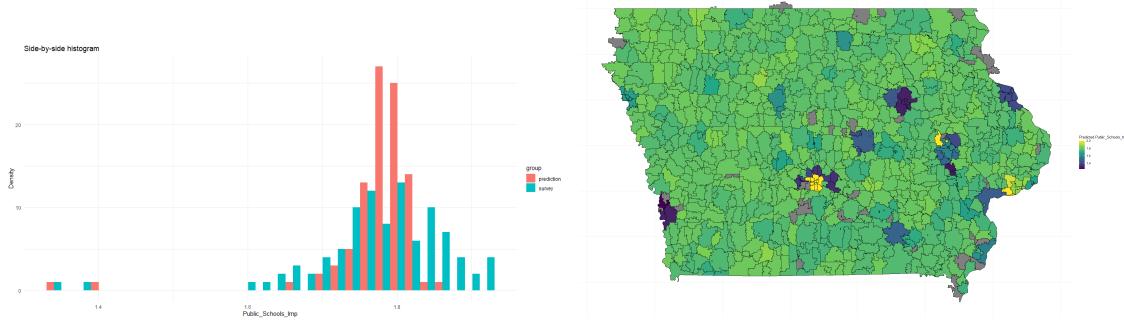


Figure 13: The distribution of predicted expected Public Schools Importance approximately follows the observed distribution.

reaching the MSE as 0.04145401. Figure 15 presents the prediction of housing satisfaction of the 99 surveyed cities based on refined ordinary linear model. And the table 2 presents the estimation of the refined ordinary linear model by removing insignificant covariates, reaching a slightly better MSE as 0.04199038.

Figure 16 presents the prediction of housing satisfaction of the 99 surveyed cities based on ordinary linear model with selected survey covariates in final random forest model. And the table 3 presents the estimation of the ordinary linear model, reaching a similar MSE as 0.04309226.

Figure 17 presents the prediction of housing satisfaction of the 99 surveyed cities based on final GAM model 11 with selected survey covariates, having MSE as 0.03421546, better than performance of ordinary linear models. And figure 20 presents the final GAM model used for prediction of housing satisfaction.

Figure 18 presents the prediction of housing satisfaction of the 99 surveyed cities based on final random forest model 4.2 with selected survey covariates, having MSE as 0.05289719, better than performance of ordinary linear models. And Figure 19 provide a comparison between final predictions of GAM model and random forest model.

In conclusion, the GAM model 11 has the best performance based on MSE.

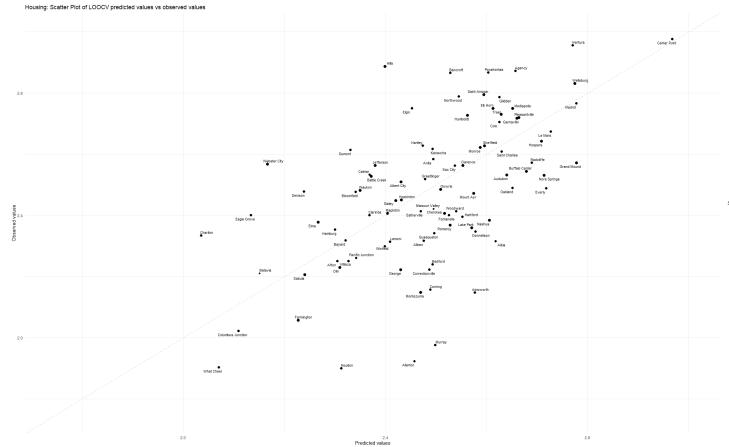


Figure 14: The surveyed cities with lower housing satisfaction is likely to be over estimated with ordinary linear model.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.5464	0.1607	9.63	0.0000
Public_Schools_Imp	-0.0449	0.0286	-1.57	0.1170
Safe	-0.1815	0.0096	-18.99	0.0000
Condition_of_Parks_Imp	0.0619	0.0355	1.74	0.0813
Unemployed	-0.0806	0.0683	-1.18	0.2380
Place_of_Residence	-0.1095	0.0163	-6.73	0.0000
higherEd_ma3	1.5820	0.2303	6.87	0.0000
unemployment_ma3	-2.9996	0.5088	-5.90	0.0000
workAdults_ma3	1.2693	0.3541	3.58	0.0003
ownerOcc_perHouse_ma3	1.2148	0.1461	8.32	0.0000
homevalueToIncome_neigh_ma3	0.1043	0.0256	4.07	0.0000

Table 1: Linear model with using selected significant survey covariates based on GAM model.

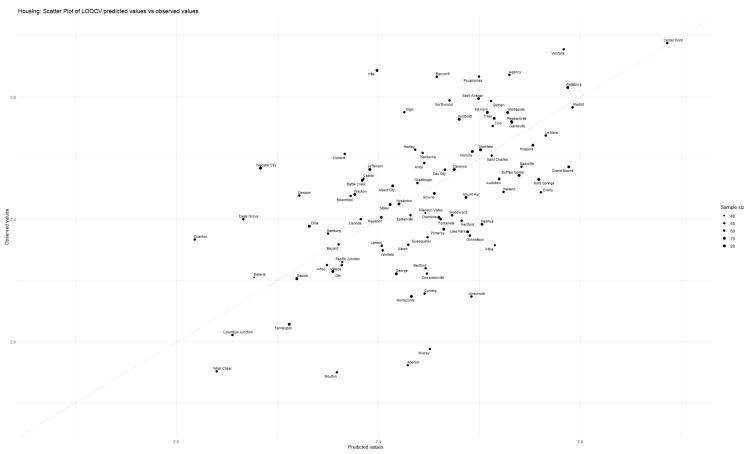


Figure 15: The surveyed cities with lower housing satisfaction is likely to be over estimated with refined ordinary linear model based on GAM model.

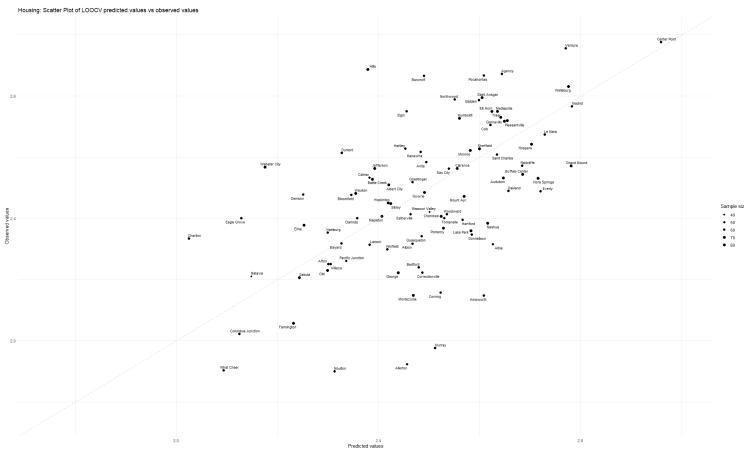


Figure 16: The surveyed cities with lower housing satisfaction is likely to be over estimated with refined ordinary linear model based on random forest.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.5824	0.1398	11.32	0.0000
Safe	-0.1826	0.0095	-19.23	0.0000
Place_of_Residence	-0.1111	0.0162	-6.85	0.0000
higherEd_ma3	1.5563	0.2300	6.77	0.0000
unemployment_ma3	-2.9790	0.5087	-5.86	0.0000
workAdults_ma3	1.2604	0.3541	3.56	0.0004
ownerOcc_perHouse_ma3	1.2191	0.1461	8.35	0.0000
homevalueToIncome_neigh_ma3	0.1064	0.0256	4.15	0.0000

Table 2: Refined linear model with using selected significant survey covariates based on GAM model.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.2909	0.1492	8.65	0.0000
Age	0.0038	0.0007	5.48	0.0000
Safe	-0.1799	0.0095	-18.98	0.0000
Place_of_Residence	-0.1036	0.0162	-6.38	0.0000
higherEd_ma3	1.5885	0.2293	6.93	0.0000
unemployment_ma3	-2.9773	0.5071	-5.87	0.0000
workAdults_ma3	1.3818	0.3537	3.91	0.0001
ownerOcc_perHouse_ma3	1.2399	0.1457	8.51	0.0000
homevalueToIncome_neigh_ma3	0.1103	0.0256	4.31	0.0000

Table 3: Ordinary linear model with using selected significant survey covariates based on random forest model.

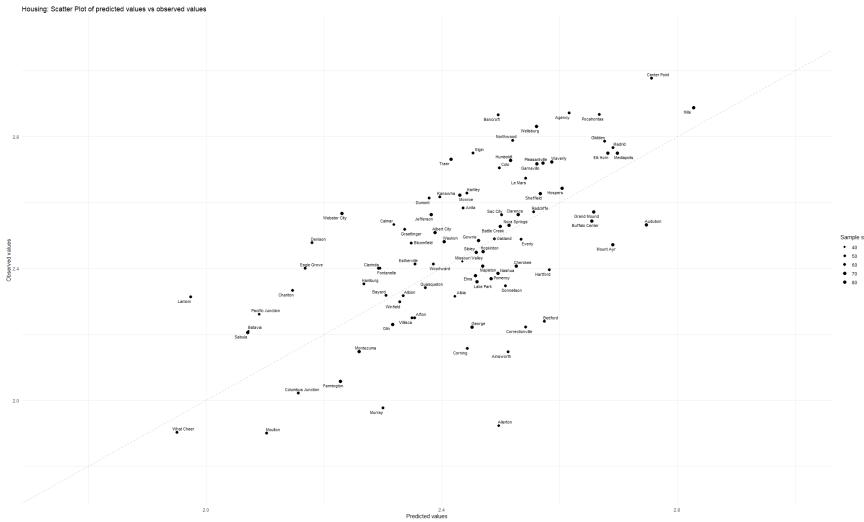


Figure 17: Approximately all surveyed cities have the housing satisfaction accurately predicted using the final GAM model.

4.5 Supplementary Models for Small Communities

Based on the common models mentioned previously, we developed supplementary, refined, models adding selected covariates limited to small communities in Iowa, corresponding to the primary ZCTA for either an urban cluster with a population between 2,500 and 10,000 or a place in a rural area with a population between 500 and 2,500.

Maintaining same survey covariates as in linear models, GAM model 11 and random forest model 9, we add extra city-level covariates, NatAmen.Natural_amenity_Scale, county.Limited access to healthy foods_% Limited Access, county.Inadequate social support_% No Social-Emotional Support,

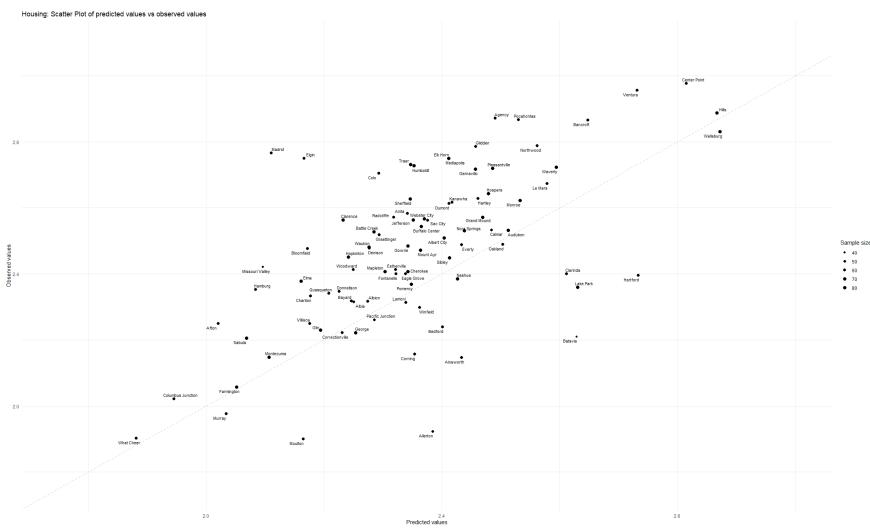


Figure 18: The housing satisfaction of the 99 surveyed cities are slightly under estimated based on the final random forest model .

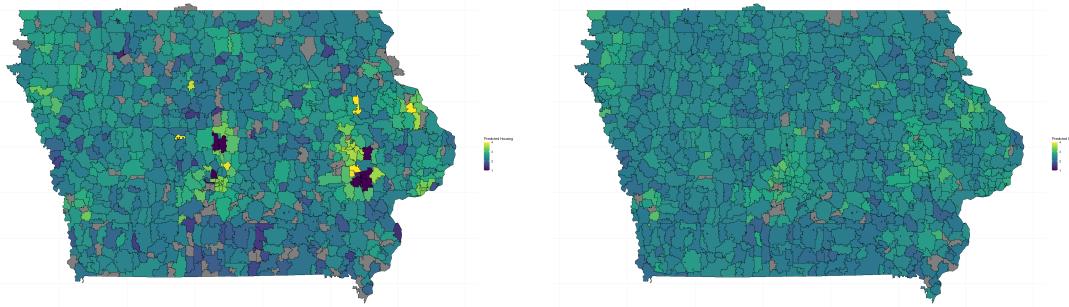


Figure 19: Prediction of housing satisfaction of cities in Iowa from GAM model (left) and random forest model (right) shows similar patterns, but predictions based on GAM model have more variability and predictions based on random forest model are more concentrated.

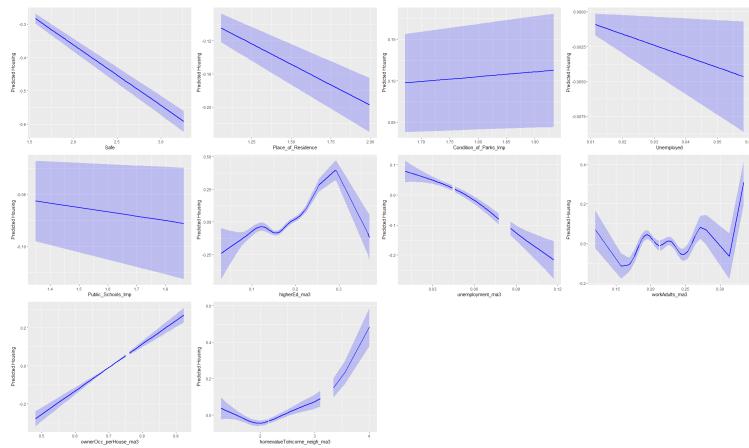


Figure 20: The final GAM model used for prediction of housing satisfaction.

`Crime.crime_rate_idx_log`, as $z_{i,k}$ for $i \in \{6, \dots, 9\}$.

Figure 21 presents the prediction of housing satisfaction of the 99 surveyed cities based on supplementary ordinary linear model with selected survey covariates in final GAM model. And the table 4 presents the estimation of the supplementary ordinary linear model, reaching the MSE as 0.03904326. Figure 22 presents the prediction of housing satisfaction of the 99 surveyed cities based on refined supplementary ordinary linear model. And the table 5 presents the estimation of the refined ordinary linear model by removing unsignificant covariates, reaching a slightly better MSE as 0.03863276.

Figure 25 presents the prediction of housing satisfaction of the 99 surveyed cities based on supplementary ordinary linear model with selected survey covariates in final random forest model. And the table 6 presents the estimation of the supplementary ordinary linear model, reaching a similar MSE as 0.03976579.

Figure 23 presents the prediction of housing satisfaction of the 99 surveyed cities based on final supplementary GAM model with selected survey covariates, having MSE as 0.01750176, better than performance of supplementary ordinary linear models. And figure 24 presents the final GAM model used for prediction of housing satisfaction.

Figure 26 presents the prediction of housing satisfaction of the 99 surveyed cities based on supplementary final random forest model with selected survey covariates, having MSE as 0.01933796, better than performance of ordinary linear models. And the housing satisfaction of the surveyed cities are slightly under estimated based on Figure 26.

In conclusion, the supplementary GAM model and random forest model have similar accuracy on prediction, and both outperform the common final models for all cities in Iowa within small communities.

Figure 27 and Figure 28 provide a comparison between final predictions of GAM model and random forest model. Table 7 presents the 10 cities that have the largest differences in absolute value between the prediction of housing satisfaction from the supplementary GAM model (the GAM model with additional variables) and the prediction from the common GAM model.

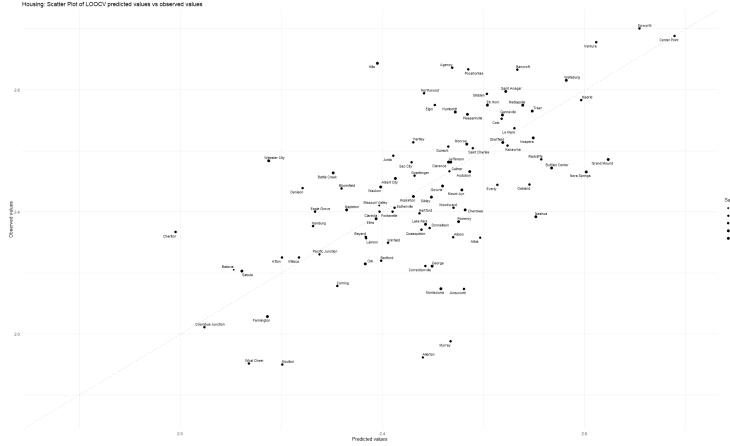


Figure 21: The surveyed cities with lower housing satisfaction is likely to be over estimated with supplementary ordinary linear model with variables selected from supplementary GAM model.

4.6 Interpretation of Supplementary GAM model

Among all the discussed models, supplementary GAM model, the Figure 24, has the highest accuracy on predicting the housing satisfaction within the 99 surveyed cities, reaching the MSE of 0.01750176. Even though the extra variables in supplementary GAM model are only available in small communities in our current processed dataset as mentioned

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.4415	0.1912	7.54	0.0000
Public_Schools_Imp	-0.0370	0.0286	-1.30	0.1950
Safe	-0.1815	0.0095	-19.06	0.0000
Condition_of_Parks_Imp	0.0636	0.0354	1.80	0.0719
Unemployed	-0.0812	0.0680	-1.19	0.2329
Place_of_Residence	-0.1032	0.0163	-6.35	0.0000
higherEd_ma3	1.5131	0.2336	6.48	0.0000
unemployment_ma3	-3.0132	0.5130	-5.87	0.0000
workAdults_ma3	1.0604	0.3558	2.98	0.0029
ownerOcc_perHouse_ma3	1.0884	0.1497	7.27	0.0000
homevalueToIncome_neigh_ma3	0.1112	0.0265	4.20	0.0000
Natural_amenity_Scale	-0.0598	0.0114	-5.23	0.0000
Limited_access_to_healthy_foods_percent_Limited_Access	0.5031	0.3539	1.42	0.1553
Inadequate_social_support_percent_No_Social_Emotional_Support	1.2575	0.3439	3.66	0.0003
crime_rate_idx_log	-0.0269	0.0100	-2.71	0.0068

Table 4: Supplementary linear model with using selected significant survey covariates based on supplementary GAM model, identifying Public_Schools_Imp, Condition_of_Parks_Imp, Unemployed, Limited_access_to_healthy_foods_%_Limited_Access as non-significant.

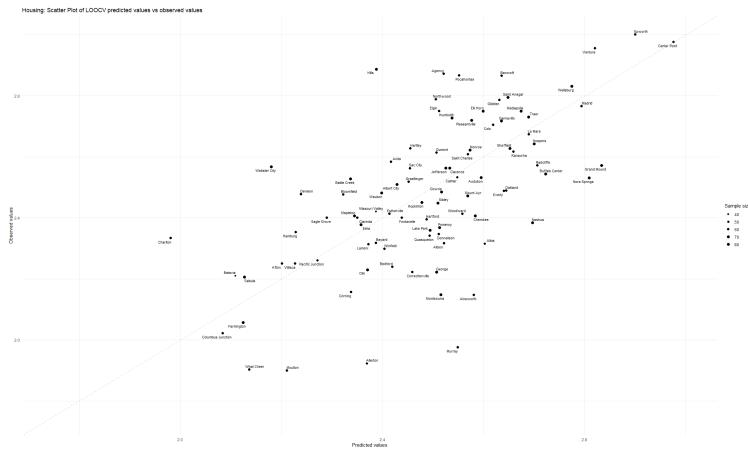


Figure 22: The surveyed cities with lower housing satisfaction is likely to be over estimated with refined supplementary ordinary linear model with variables selected from supplementary GAM model.

in section 3.5, the observations in the original dataset should be recorded on county-level based on description from the sources. This means we can actually extend the prediction to all the cities in Iowa, instead of only small communities, which requires some extra effort on data processing.

The final supplementary GAM, the Figure 24, relates housing satisfaction to two sets of predictors: (i) respondent-level survey covariates that enter linearly (constant-slope effects) and (ii) city-level covariates that enter as smooth functions, allowing for non-linear partial effects. We read signs and magnitudes for linear terms as ceteris paribus changes in the expected outcome, while smooth terms are interpreted from their partial-effect curves (with pointwise confidence bands), where upward (downward) segments indicate higher (lower) expected satisfaction as the covariate increases.

Survey covariates (linear effects)

- **Safe** (1 = “safe”, 7 = “dangerous”): Estimated slope is negative; higher (less-safe) values are associated with lower expected housing satisfaction.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.5718	0.1655	9.50	0.0000
Safe	-0.1827	0.0095	-19.31	0.0000
Place_of_Residence	-0.1054	0.0162	-6.51	0.0000
higherEd_ma3	1.4367	0.2305	6.23	0.0000
unemployment_ma3	-3.0091	0.5128	-5.87	0.0000
workAdults_ma3	1.0769	0.3554	3.03	0.0025
ownerOcc_perHouse_ma3	1.0594	0.1481	7.15	0.0000
homevalueToIncome_neigh_ma3	0.1048	0.0258	4.07	0.0000
Natural_amenity_Scale	-0.0589	0.0114	-5.17	0.0000
Inadequate_social_support_percent_No_Social_Emotional_Support	1.2731	0.3436	3.71	0.0002
crime_rate_idx_log	-0.0271	0.0100	-2.72	0.0065

Table 5: Refined supplementary linear model with using selected significant survey covariates based on Supplementary GAM model.

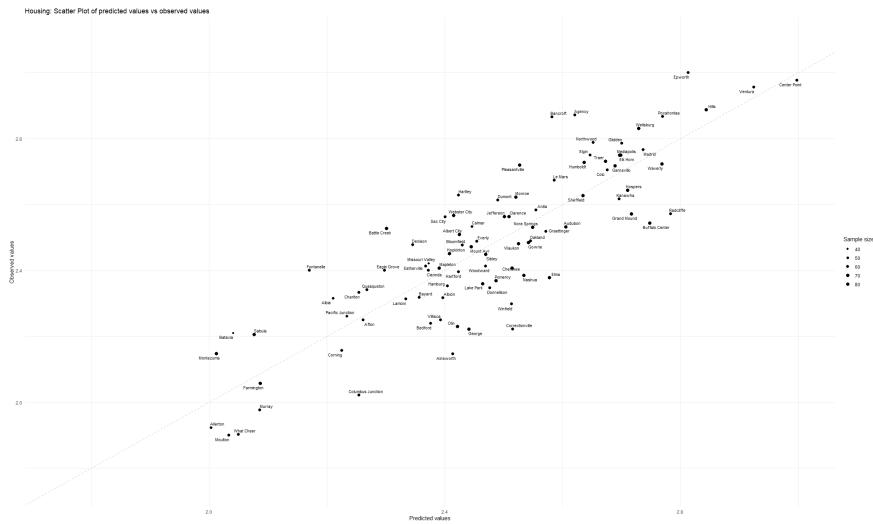


Figure 23: Approximately all surveyed cities have the housing satisfaction accurately predicted using the final supplementary GAM model.

- **Place of Residence** (1 = within city; 2 = outside on a farm; 3 = outside, not on a farm): Negative slope; being outside city limits is associated with lower expected satisfaction.
- **Condition of Parks** (No/Yes): Small positive linear effect; respondents emphasizing park condition show slightly higher expected satisfaction.
- **Unemployed** (0/1 at respondent): Small negative effect; unemployment is associated with lower expected satisfaction.
- **Public Schools** (No/Yes): Very small negative-to-near-zero effect once city-level covariates are controlled.

City-level covariates (smooth terms)

- **acs.higherEd (ma3)** (share age 25+ with \geq BA): Partial effect increases with the variable; higher educational attainment corresponds to higher expected satisfaction, with diminishing returns at the upper end.
- **acs.unemployment (ma3)**: Decreasing partial effect; higher unemployment is associated with lower expected satisfaction.

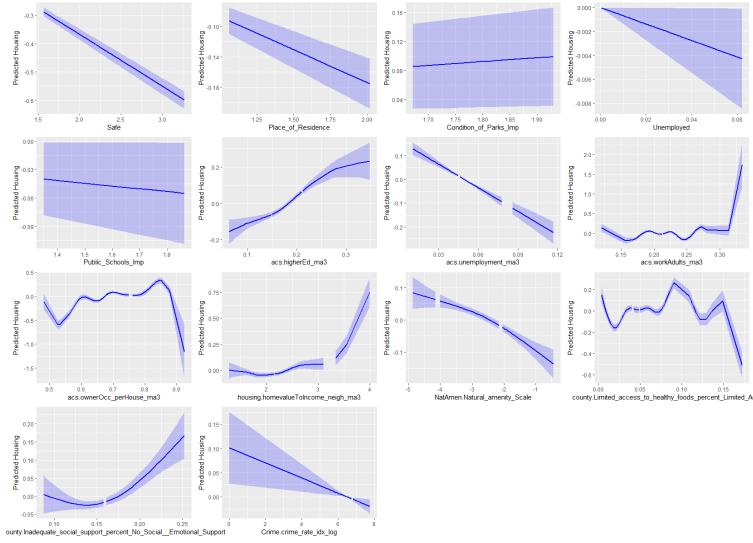


Figure 24: The final supplementary GAM model used for prediction of housing satisfaction.

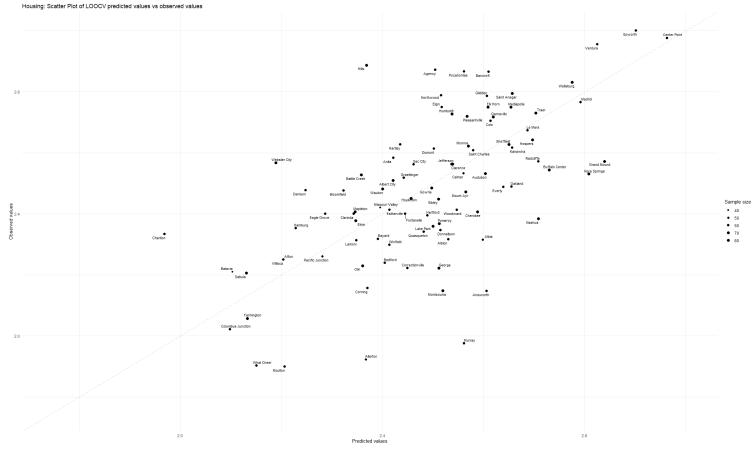


Figure 25: The surveyed cities with lower housing satisfaction is likely to be over estimated with refined ordinary linear model based on variables selected from supplementary random forest model.

- **acs.workAdults (ma3)** (share working-age/working): Increasing partial effect; more working adults correspond to higher expected satisfaction.
- **acs.ownerOcc perHouse (ma3)**: Increasing, roughly linear partial effect; higher owner-occupancy predicts higher satisfaction.
- **housing.homevalueToIncome_neigh (ma3)**: Mild increasing partial effect; higher home-value-to-income (neighborhood-level) is associated with slightly higher satisfaction.
- **NatAmen. Natural Amenity Scale**: Decreasing partial effect in the small-communities subset; higher amenity score is linked to lower satisfaction in this sample.
- **County: Limited access to healthy foods (% Limited Access)**: Near-flat to weakly negative partial effect; higher limited access tends to reduce satisfaction, but the effect is small.
- **County: Inadequate social support (% no social/emotional support)**: Increasing partial effect in this subset; interpret cautiously due to colocation with other place features.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.2740	0.1735	7.34	0.0000
Age	0.0038	0.0007	5.52	0.0000
Safe	-0.1800	0.0094	-19.06	0.0000
Place_of_Residence	-0.0978	0.0162	-6.04	0.0000
higherEd_ma3	1.4689	0.2299	6.39	0.0000
unemployment_ma3	-3.0032	0.5112	-5.87	0.0000
workAdults_ma3	1.1968	0.3550	3.37	0.0008
ownerOcc_perHouse_ma3	1.0795	0.1477	7.31	0.0000
homevalueToIncome_neigh_ma3	0.1088	0.0257	4.23	0.0000
Natural_amenity_Scale	-0.0592	0.0114	-5.21	0.0000
Inadequate_social_support_percent_No_Social_Emotional_Support	1.2871	0.3425	3.76	0.0002
crime_rate_idx_log	-0.0267	0.0099	-2.69	0.0071

Table 6: Supplementary ordinary linear model with using selected significant survey covariates based on supplementary random forest model, where Limited access to healthy foods % Limited Access is identified as non-significant.

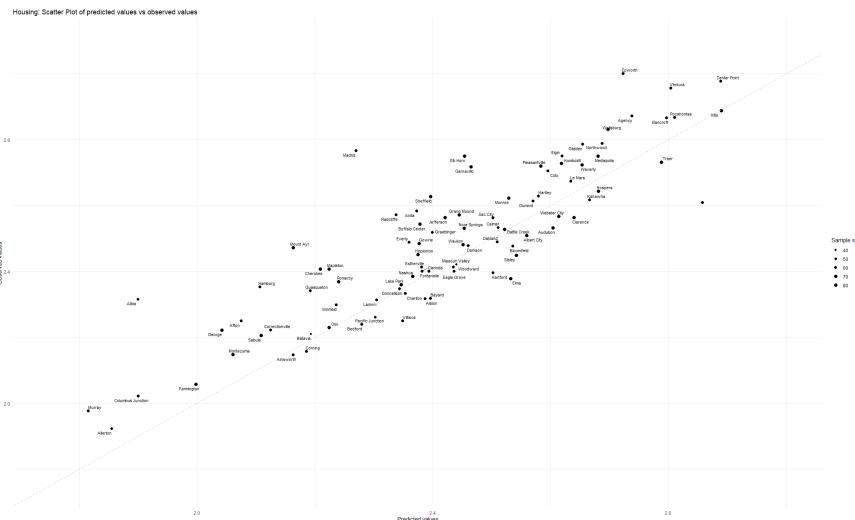


Figure 26: The housing satisfaction of the 99 surveyed cities are slightly under estimated based on the final supplementary random forest model .

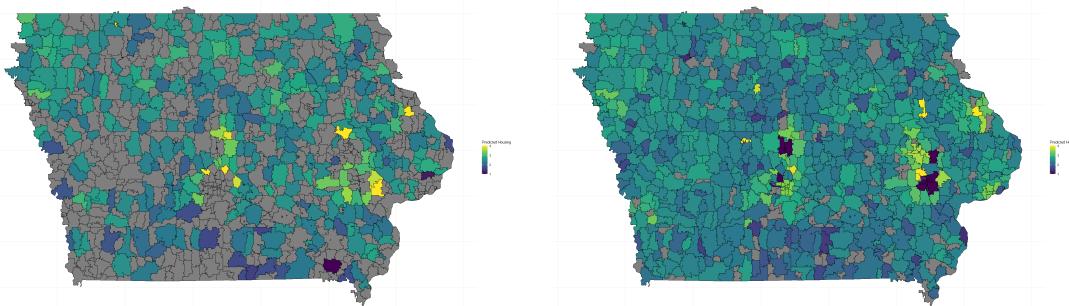


Figure 27: Prediction of housing satisfaction of cities in Iowa from supplementary GAM model (left) and common GAM model (right) shows similar patterns, where areas closer to larger cities corresponds to higher average housing satisfaction.

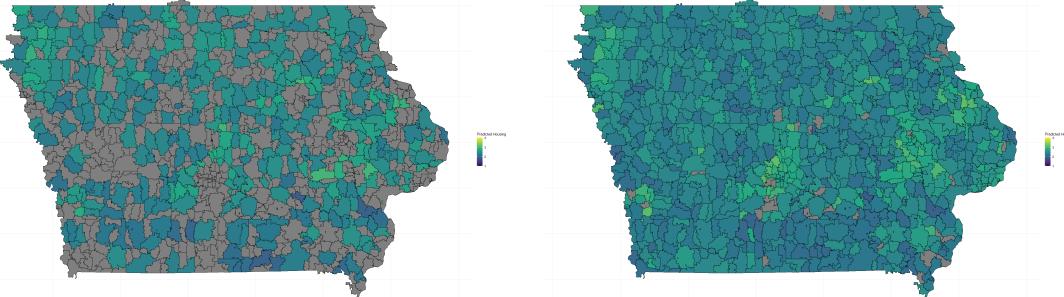


Figure 28: Prediction of housing satisfaction of cities in Iowa from supplementary random forest model (left) and common random forest model (right) shows similar patterns, where areas closer to larger cities corresponds to higher average housing satisfaction..

- **Crime: crime rate index (log):** Decreasing, roughly linear partial effect; higher crime corresponds to lower expected satisfaction.

place	common GAM prediction	supplementary GAM prediction	difference
Okoboji	1.41	4.00	2.59
Mount Vernon	1.00	3.14	2.14
Orleans	1.64	3.56	1.92
Arnolds Park	2.18	4.00	1.82
West Liberty	2.40	4.00	1.60
Keosauqua	2.38	1.00	1.38
Mitchellville	2.57	3.77	1.20
Long Grove	2.38	1.19	1.19
Norway	2.73	1.79	0.94
Solon	2.24	3.00	0.76

Table 7: The 10 cities from Figure 27 that have the largest differences in absolute value between the prediction of common GAM model and the prediction of supplementary GAM model

5 References

References

- [1] R. Batista, Z. Zhu, D. Peters, and K. Zarecor, “Predicting resident satisfaction with public schools in small town iowa,” *Stat*, vol. 12, no. 1, p. e517, 2023.
- [2] S. N. Wood, “Inference and computation with generalized additive models and their extensions,” *Test*, vol. 29, no. 2, pp. 307–339, 2020.
- [3] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, “Vsurf: an r package for variable selection using random forests,” *The R journal*, vol. 7, no. 2, pp. 19–33, 2015.
- [4] R. Diaz-Uriarte and S. A. de Andrés, “Variable selection from random forests: application to gene expression data,” *arXiv preprint q-bio/0503025*, 2005.
- [5] M. N. Wright and A. Ziegler, “ranger: A fast implementation of random forests for high dimensional data in c++ and r,” *Journal of statistical software*, vol. 77, pp. 1–17, 2017.