Daniel Prusinski

412

Appendix 1 for Assignment 6 – Geo Spatial Data Modeling

Output 1: Metadata

print(str(houses))

'data.frame': 20640 obs. of 9 variables:

$ value : num 452600 358500 352100 341300 342200 ...

$ income : num 8.33 8.3 7.26 5.64 3.85 ...

$ age : num 41 21 52 52 52 52 52 52 42 52 ...

$ rooms : num 880 7099 1467 1274 1627 ...

$ bedrooms : num 129 1106 190 235 280 ...

$ pop : num 322 2401 496 558 565 ...

$ hh : num 126 1138 177 219 259 ...

$ latitude : num 37.9 37.9 37.9 37.9 37.9 ...

$ longitude: num -122 -122 -122 -122 -122 ...

NULL

Output 2: Distribution of Data

value income age rooms

Min. : 14999 Min. : 0.4999 Min. : 1.00 Min. : 2

1st Qu.:119600 1st Qu.: 2.5634 1st Qu.:18.00 1st Qu.: 1448

Median :179700 Median : 3.5348 Median :29.00 Median : 2127

Mean :206856 Mean : 3.8707 Mean :28.64 Mean : 2636

3rd Qu.:264725 3rd Qu.: 4.7432 3rd Qu.:37.00 3rd Qu.: 3148

Max. :500001 Max. :15.0001 Max. :52.00 Max. :39320

bedrooms pop hh latitude

Min. : 1.0 Min. : 3 Min. : 1.0 Min. :32.54

1st Qu.: 295.0 1st Qu.: 787 1st Qu.: 280.0 1st Qu.:33.93

Median : 435.0 Median : 1166 Median : 409.0 Median :34.26

Mean : 537.9 Mean : 1425 Mean : 499.5 Mean :35.63

3rd Qu.: 647.0 3rd Qu.: 1725 3rd Qu.: 605.0 3rd Qu.:37.71

Max. :6445.0 Max. :35682 Max. :6082.0 Max. :41.95

longitude

Min. :-124.3

1st Qu.:-121.8

Median :-118.5

Mean :-119.6

3rd Qu.:-118.0

Max. :-114.3

vru\_time q\_start q\_exit q\_time

Min. :-192.00 Length:33344 Length:33344 Min. : 0.00

1st Qu.: 6.00 Class :character Class :character 1st Qu.: 0.00

Median : 9.00 Mode :character Mode :character Median : 9.00

Mean : 10.46 Mean : 41.79

3rd Qu.: 11.00 3rd Qu.: 57.00

Max. :1860.00 Max. :908.00

outcome ser\_start ser\_exit ser\_time

AGENT :27162 Length:33344 Length:33344 Min. : 0.0

HANG : 5904 Class :character Class :character 1st Qu.: 15.0

PHANTOM: 278 Mode :character Mode :character Median : 80.0

Mean : 144.3

3rd Qu.: 180.0

Max. :4264.0

server

Length:33344

Class :character

Mode :character





Output 3: Variables after Log Transformations

print(str(houses))

'data.frame': 20640 obs. of 17 variables:

$ value : num 452600 358500 352100 341300 342200 ...

$ income : num 8.33 8.3 7.26 5.64 3.85 ...

$ age : num 41 21 52 52 52 52 52 52 42 52 ...

$ rooms : num 880 7099 1467 1274 1627 ...

$ bedrooms : num 129 1106 190 235 280 ...

$ pop : num 322 2401 496 558 565 ...

$ hh : num 126 1138 177 219 259 ...

$ latitude : num 37.9 37.9 37.9 37.9 37.9 ...

$ longitude : num -122 -122 -122 -122 -122 ...

$ log\_value : num 13 12.8 12.8 12.7 12.7 ...

$ income\_squared : num 69.3 68.9 52.7 31.8 14.8 ...

$ income\_cubed : num 577 572.1 382.2 179.7 56.9 ...

$ log\_age : num 3.71 3.04 3.95 3.95 3.95 ...

$ log\_pc\_rooms : num 1.005 1.084 1.084 0.826 1.058 ...

$ log\_pc\_bedrooms: num -0.915 -0.775 -0.96 -0.865 -0.702 ...

$ log\_pop\_hh : num 0.938 0.747 1.03 0.935 0.78 ...

$ log\_hh : num 4.84 7.04 5.18 5.39 5.56 ...

NULL

>

> # check data frame object and variable values

> print(summary(houses))

value income age rooms

Min. : 14999 Min. : 0.4999 Min. : 1.00 Min. : 2

1st Qu.:119600 1st Qu.: 2.5634 1st Qu.:18.00 1st Qu.: 1448

Median :179700 Median : 3.5348 Median :29.00 Median : 2127

Mean :206856 Mean : 3.8707 Mean :28.64 Mean : 2636

3rd Qu.:264725 3rd Qu.: 4.7432 3rd Qu.:37.00 3rd Qu.: 3148

Max. :500001 Max. :15.0001 Max. :52.00 Max. :39320

bedrooms pop hh latitude

Min. : 1.0 Min. : 3 Min. : 1.0 Min. :32.54

1st Qu.: 295.0 1st Qu.: 787 1st Qu.: 280.0 1st Qu.:33.93

Median : 435.0 Median : 1166 Median : 409.0 Median :34.26

Mean : 537.9 Mean : 1425 Mean : 499.5 Mean :35.63

3rd Qu.: 647.0 3rd Qu.: 1725 3rd Qu.: 605.0 3rd Qu.:37.71

Max. :6445.0 Max. :35682 Max. :6082.0 Max. :41.95

longitude log\_value income\_squared income\_cubed

Min. :-124.3 Min. : 9.616 Min. : 0.2499 Min. : 0.125

1st Qu.:-121.8 1st Qu.:11.692 1st Qu.: 6.5710 1st Qu.: 16.844

Median :-118.5 Median :12.099 Median : 12.4948 Median : 44.167

Mean :-119.6 Mean :12.085 Mean : 18.5912 Mean : 111.190

3rd Qu.:-118.0 3rd Qu.:12.486 3rd Qu.: 22.4984 3rd Qu.: 106.716

Max. :-114.3 Max. :13.122 Max. :225.0030 Max. :3375.068

log\_age log\_pc\_rooms log\_pc\_bedrooms log\_pop\_hh

Min. :0.000 Min. :-5.9729 Min. :-7.3079 Min. :-0.3677

1st Qu.:2.890 1st Qu.: 0.4203 1st Qu.:-1.1531 1st Qu.: 0.8878

Median :3.367 Median : 0.6616 Median :-0.9888 Median : 1.0361

Mean :3.225 Mean : 0.6045 Mean :-0.9731 Mean : 1.0433

3rd Qu.:3.611 3rd Qu.: 0.8312 3rd Qu.:-0.8150 3rd Qu.: 1.1885

Max. :3.951 Max. : 4.0114 Max. : 2.6529 Max. : 7.1256

log\_hh

Min. :0.000

1st Qu.:5.635

Median :6.014

Mean :5.981

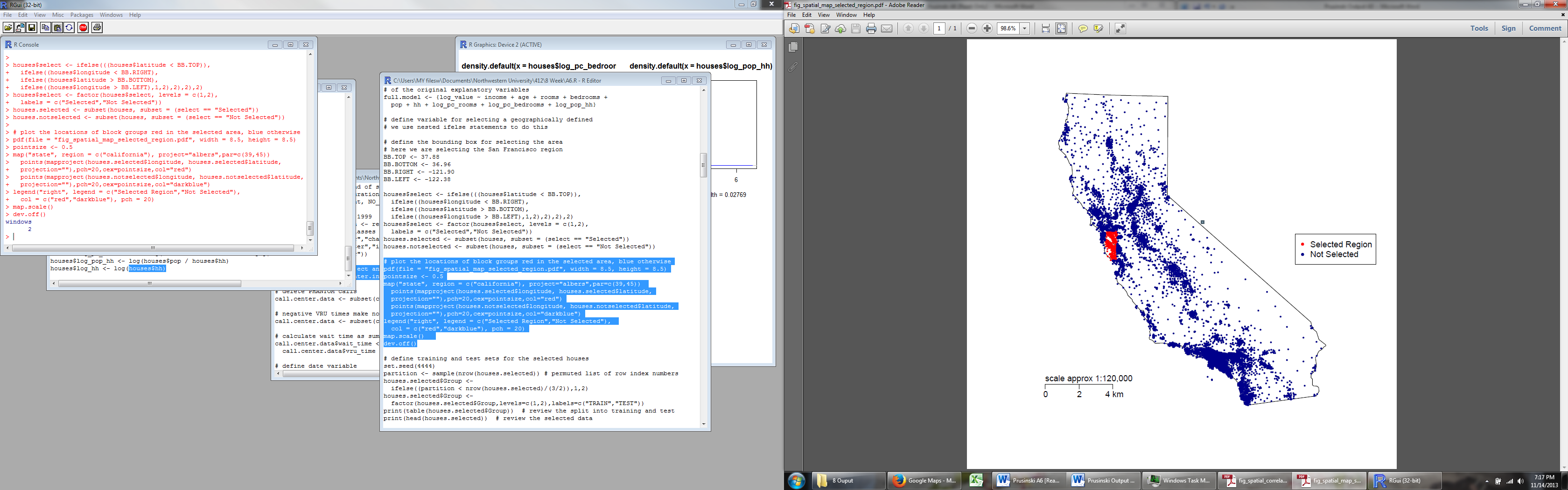
3rd Qu.:6.405

Max. :8.713





Output 4: Spatial Map Selected Region



Output 6: Correlation Map

Output 7: Linear Regression without Spatial Points

Call:

lm(formula = pace.barry.model, data = houses.train)

Coefficients:

(Intercept) income income\_squared income\_cubed

11.4237051 0.3051386 -0.0063915 -0.0003248

log\_age log\_pc\_rooms log\_pc\_bedrooms log\_pop\_hh

0.0573435 -0.3002109 0.0402894 -0.5801412

log\_hh

0.0687831

Call:

lm(formula = pace.barry.model, data = houses.train)

Residuals:

Min 1Q Median 3Q Max

-1.83104 -0.15433 -0.01033 0.16552 1.61647

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 11.4237051 0.1401238 81.526 < 2e-16 \*\*\*

income 0.3051386 0.0306469 9.957 < 2e-16 \*\*\*

income\_squared -0.0063915 0.0046487 -1.375 0.16940

income\_cubed -0.0003248 0.0002113 -1.537 0.12449

log\_age 0.0573435 0.0182177 3.148 0.00168 \*\*

log\_pc\_rooms -0.3002109 0.0572792 -5.241 1.86e-07 \*\*\*

log\_pc\_bedrooms 0.0402894 0.1096682 0.367 0.71340

log\_pop\_hh -0.5801412 0.1067262 -5.436 6.51e-08 \*\*\*

log\_hh 0.0687831 0.0122027 5.637 2.12e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2842 on 1298 degrees of freedom

Multiple R-squared: 0.6452, Adjusted R-squared: 0.643

F-statistic: 295 on 8 and 1298 DF, p-value: < 2.2e-16

>

> # direct calculation of root-mean-squared prediction error

[1] 0.2832567

> # report R-squared on training data

> print(cor(houses.train$log\_value,predict(pace.barry.train.fit))^2)

[1] 0.6451926

Training set proportion of variance accounted for by linear regression = 0.645>

> # test model fit to training set on the test set

[1] 0.2584695

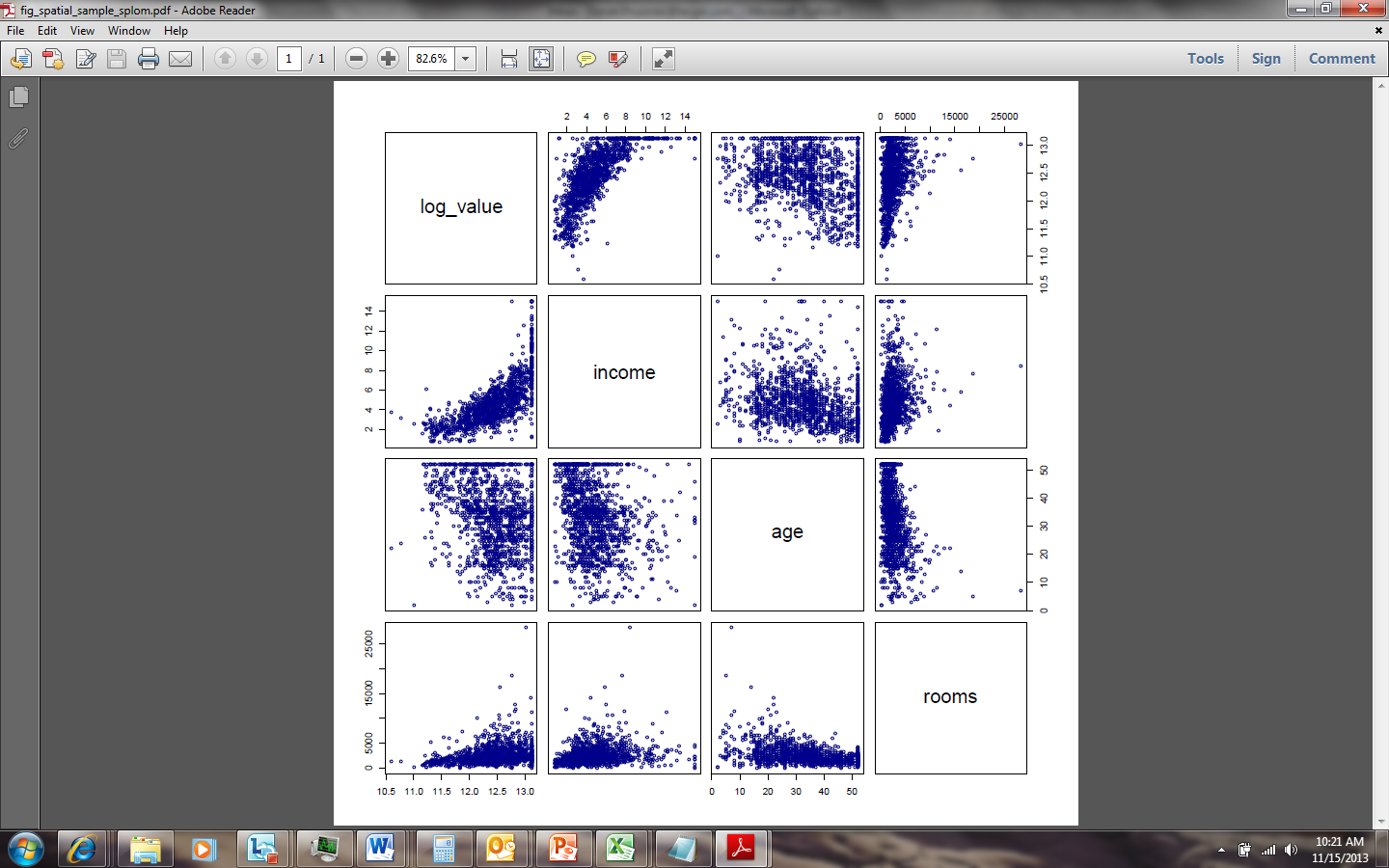
[1] 0.7015261

>

10-fold CV results:

CV

0.1961384



Output 8: Full Tree Regression

CP nsplit rel error xerror xstd

1 0.39438424 0 1.0000000 1.0021816 0.03788014

2 0.10644598 1 0.6056158 0.6312032 0.02541968

3 0.06723167 2 0.4991698 0.5225227 0.02540437

4 0.02706050 3 0.4319381 0.4539892 0.02397689

5 0.02457305 4 0.4048776 0.4449655 0.02520428

6 0.01701520 5 0.3803046 0.4112027 0.02496583

7 0.01634933 6 0.3632894 0.4047483 0.02489414

8 0.01434153 7 0.3469400 0.4016213 0.02495083

9 0.01000000 8 0.3325985 0.3964548 0.02431400

Variable importance

income log\_pc\_rooms log\_pc\_bedrooms log\_pop\_hh age

43 18 12 11 6

rooms hh bedrooms pop

6 1 1 1

Node number 1: 1307 observations, complexity param=0.3943842

mean=12.42984, MSE=0.2261351

left son=2 (711 obs) right son=3 (596 obs)

Primary splits:

income < 4.62305 to the left, improve=0.39438420, (0 missing)

log\_pc\_rooms < 0.708089 to the left, improve=0.26342530, (0 missing)

rooms < 1941 to the left, improve=0.10513490, (0 missing)

age < 35.5 to the right, improve=0.09750198, (0 missing)

log\_pop\_hh < 1.164632 to the right, improve=0.05747476, (0 missing)

Surrogate splits:

log\_pc\_rooms < 0.8192164 to the left, agree=0.721, adj=0.388, (0 split)

log\_pc\_bedrooms < -0.8752156 to the right, agree=0.651, adj=0.235, (0 split)

age < 35.5 to the right, agree=0.630, adj=0.190, (0 split)

rooms < 2705.5 to the left, agree=0.627, adj=0.181, (0 split)

log\_pop\_hh < 0.8740992 to the left, agree=0.617, adj=0.159, (0 split)

Node number 2: 711 observations, complexity param=0.106446

mean=12.15641, MSE=0.1779834

left son=4 (215 obs) right son=5 (496 obs)

Primary splits:

income < 2.66755 to the left, improve=0.2486128, (0 missing)

log\_pop\_hh < 0.9774789 to the right, improve=0.1569347, (0 missing)

log\_pc\_rooms < 0.532225 to the left, improve=0.1198823, (0 missing)

rooms < 1806 to the left, improve=0.1180630, (0 missing)

hh < 420.5 to the left, improve=0.1147604, (0 missing)

Surrogate splits:

log\_pc\_rooms < 0.03791093 to the left, agree=0.716, adj=0.060, (0 split)

rooms < 405 to the left, agree=0.710, adj=0.042, (0 split)

hh < 126.5 to the left, agree=0.710, adj=0.042, (0 split)

pop < 218.5 to the left, agree=0.709, adj=0.037, (0 split)

log\_pc\_bedrooms < -1.666657 to the left, agree=0.707, adj=0.033, (0 split)

Node number 3: 596 observations, complexity param=0.06723167

mean=12.75601, MSE=0.08800114

left son=6 (343 obs) right son=7 (253 obs)

Primary splits:

income < 6.33795 to the left, improve=0.37886360, (0 missing)

log\_pc\_rooms < 0.78489 to the left, improve=0.28082770, (0 missing)

log\_pop\_hh < 1.174025 to the right, improve=0.07793179, (0 missing)

log\_pc\_bedrooms < -1.074462 to the left, improve=0.06883916, (0 missing)

pop < 1621.5 to the right, improve=0.02589955, (0 missing)

Surrogate splits:

log\_pc\_rooms < 0.914585 to the left, agree=0.772, adj=0.462, (0 split)

hh < 217 to the right, agree=0.601, adj=0.059, (0 split)

bedrooms < 223.5 to the right, agree=0.599, adj=0.055, (0 split)

pop < 582 to the right, agree=0.591, adj=0.036, (0 split)

log\_pc\_bedrooms < -0.9242346 to the right, agree=0.591, adj=0.036, (0 split)

Node number 4: 215 observations, complexity param=0.02457305

mean=11.83691, MSE=0.165809

left son=8 (144 obs) right son=9 (71 obs)

Primary splits:

log\_pop\_hh < 0.8187751 to the right, improve=0.2037304, (0 missing)

log\_pc\_bedrooms < -0.7643257 to the left, improve=0.1634555, (0 missing)

hh < 408 to the left, improve=0.1367841, (0 missing)

bedrooms < 433.5 to the left, improve=0.1195585, (0 missing)

pop < 1190 to the left, improve=0.1101065, (0 missing)

Surrogate splits:

log\_pc\_bedrooms < -0.7558615 to the left, agree=0.921, adj=0.761, (0 split)

log\_pc\_rooms < 0.7259301 to the left, agree=0.772, adj=0.310, (0 split)

hh < 520 to the left, agree=0.735, adj=0.197, (0 split)

rooms < 2809 to the left, agree=0.726, adj=0.169, (0 split)

bedrooms < 649.5 to the left, agree=0.726, adj=0.169, (0 split)

Node number 5: 496 observations, complexity param=0.0270605

mean=12.29491, MSE=0.1198312

left son=10 (274 obs) right son=11 (222 obs)

Primary splits:

log\_pop\_hh < 0.8707028 to the right, improve=0.13456370, (0 missing)

log\_pc\_bedrooms < -0.8947055 to the left, improve=0.11769040, (0 missing)

rooms < 2019.5 to the left, improve=0.08868108, (0 missing)

hh < 423.5 to the left, improve=0.08866103, (0 missing)

bedrooms < 409.5 to the left, improve=0.08822457, (0 missing)

Surrogate splits:

log\_pc\_bedrooms < -0.8381954 to the left, agree=0.915, adj=0.811, (0 split)

log\_pc\_rooms < 0.6579236 to the left, agree=0.726, adj=0.387, (0 split)

hh < 665 to the left, agree=0.637, adj=0.189, (0 split)

bedrooms < 687.5 to the left, agree=0.635, adj=0.185, (0 split)

age < 30.5 to the right, agree=0.631, adj=0.176, (0 split)

Node number 6: 343 observations, complexity param=0.01434153

mean=12.5992, MSE=0.07010273

left son=12 (185 obs) right son=13 (158 obs)

Primary splits:

log\_pop\_hh < 0.9635097 to the right, improve=0.17628270, (0 missing)

log\_pc\_rooms < 0.7706807 to the left, improve=0.16965260, (0 missing)

log\_pc\_bedrooms < -1.084654 to the left, improve=0.15027320, (0 missing)

income < 5.37135 to the left, improve=0.05541173, (0 missing)

age < 49.5 to the left, improve=0.03725613, (0 missing)

Surrogate splits:

log\_pc\_bedrooms < -0.9449526 to the left, agree=0.895, adj=0.772, (0 split)

log\_pc\_rooms < 0.7620612 to the left, agree=0.770, adj=0.500, (0 split)

age < 38.5 to the left, agree=0.659, adj=0.259, (0 split)

pop < 1103 to the right, agree=0.586, adj=0.101, (0 split)

bedrooms < 1011.5 to the left, agree=0.569, adj=0.063, (0 split)

Node number 7: 253 observations

mean=12.96862, MSE=0.03372549

Node number 8: 144 observations, complexity param=0.01634933

mean=11.70786, MSE=0.1341078

left son=16 (131 obs) right son=17 (13 obs)

Primary splits:

log\_pop\_hh < 1.431683 to the left, improve=0.25022280, (0 missing)

log\_pc\_bedrooms < -1.397417 to the right, improve=0.24649740, (0 missing)

log\_pc\_rooms < -0.2011101 to the right, improve=0.24010050, (0 missing)

pop < 1190 to the left, improve=0.13455880, (0 missing)

hh < 406.5 to the left, improve=0.09877935, (0 missing)

Surrogate splits:

log\_pc\_bedrooms < -1.412594 to the right, agree=0.986, adj=0.846, (0 split)

log\_pc\_rooms < -0.350903 to the right, agree=0.958, adj=0.538, (0 split)

rooms < 100.5 to the right, agree=0.924, adj=0.154, (0 split)

bedrooms < 36.5 to the right, agree=0.917, adj=0.077, (0 split)

hh < 36.5 to the right, agree=0.917, adj=0.077, (0 split)

Node number 9: 71 observations

mean=12.09866, MSE=0.1278119

Node number 10: 274 observations

mean=12.18061, MSE=0.08968683

Node number 11: 222 observations, complexity param=0.0170152

mean=12.43598, MSE=0.1210095

left son=22 (7 obs) right son=23 (215 obs)

Primary splits:

log\_pc\_rooms < 1.068216 to the right, improve=0.18720100, (0 missing)

income < 3.31965 to the left, improve=0.11992980, (0 missing)

log\_pc\_bedrooms < -0.3974216 to the right, improve=0.07871486, (0 missing)

log\_pop\_hh < 0.4496501 to the left, improve=0.07760785, (0 missing)

pop < 717.5 to the left, improve=0.05319119, (0 missing)

Node number 12: 185 observations

mean=12.49646, MSE=0.05628792

Node number 13: 158 observations

mean=12.71949, MSE=0.0594507

Node number 16: 131 observations

mean=11.65015, MSE=0.08620894

Node number 17: 13 observations

mean=12.28936, MSE=0.2450745

Node number 22: 7 observations

mean=11.60185, MSE=0.4893609

Node number 23: 215 observations

mean=12.46314, MSE=0.08562601

n= 1307

node), split, n, deviance, yval

\* denotes terminal node

1) root 1307 295.558500 12.42984

2) income< 4.62305 711 126.546200 12.15641

4) income< 2.66755 215 35.648940 11.83691

8) log\_pop\_hh>=0.8187751 144 19.311520 11.70786

16) log\_pop\_hh< 1.431683 131 11.293370 11.65015 \*

17) log\_pop\_hh>=1.431683 13 3.185969 12.28936 \*

9) log\_pop\_hh< 0.8187751 71 9.074646 12.09866 \*

5) income>=2.66755 496 59.436260 12.29491

10) log\_pop\_hh>=0.8707028 274 24.574190 12.18061 \*

11) log\_pop\_hh< 0.8707028 222 26.864100 12.43598

22) log\_pc\_rooms>=1.068216 7 3.425526 11.60185 \*

23) log\_pc\_rooms< 1.068216 215 18.409590 12.46314 \*

3) income>=4.62305 596 52.448680 12.75601

6) income< 6.33795 343 24.045240 12.59920

12) log\_pop\_hh>=0.9635097 185 10.413270 12.49646 \*

13) log\_pop\_hh< 0.9635097 158 9.393211 12.71949 \*

7) income>=6.33795 253 8.532549 12.96862 \*

>

> # root-mean-squared for trees on training set

[1] 0.2742484

> # report R-squared on training data

[1] 0.6674015

Training set proportion of variance accounted for by tree-structured regression (full model) = 0.667>

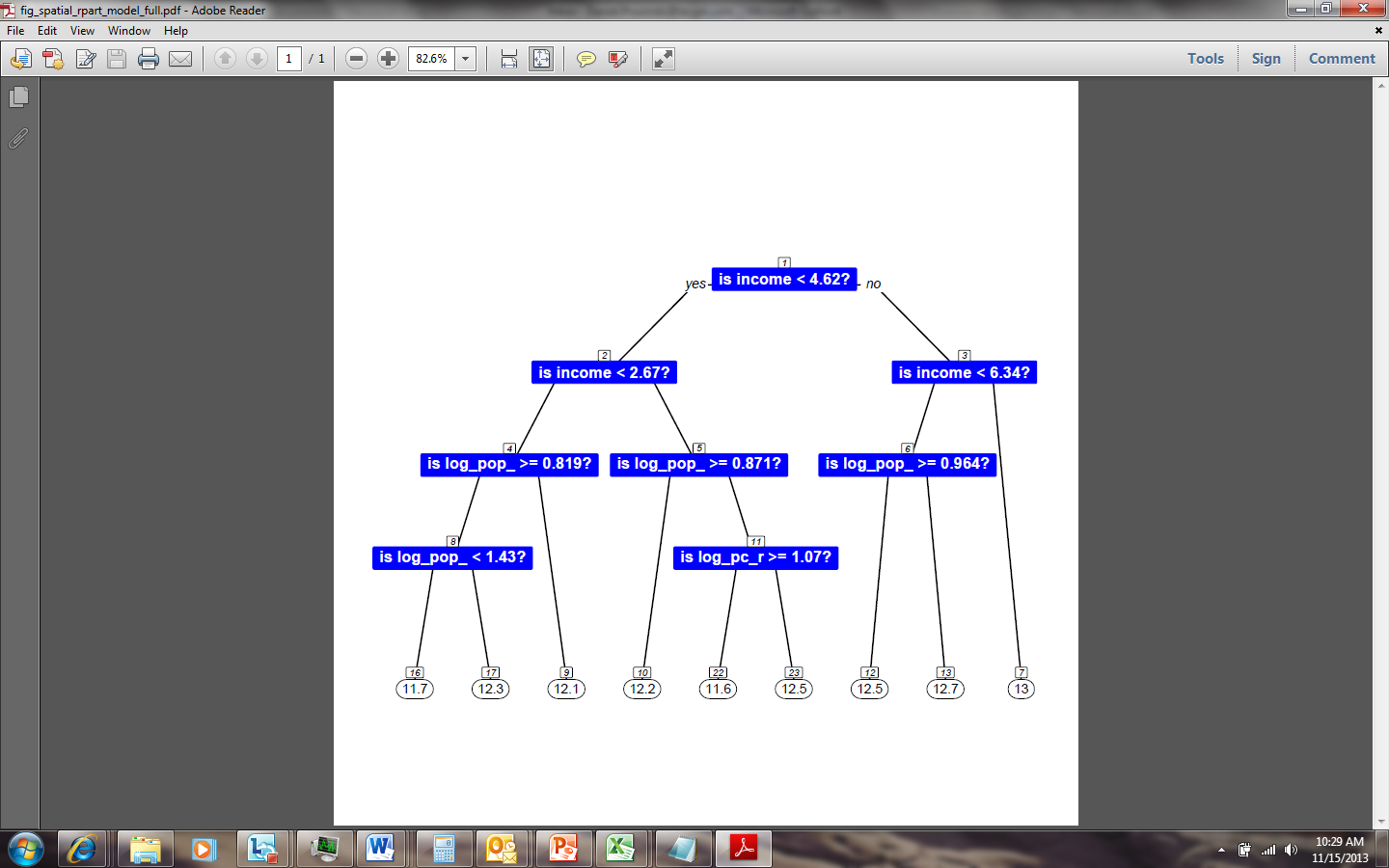
> # root-mean-squared for trees on test set

[1] 0.2936755

> # report R-squared on training data

> print(cor(houses.test$log\_value,houses.test$rpart.train.fit.full.pred)^2)

[1] 0.6191817

> 

Output 9: Random Forest

set.seed (9999)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 3

Mean of squared residuals: 0.06922999

% Var explained: 69.39

root-mean-squared for random forest on training set

[1] 0.1156064

R-squared on training data

[1] 0.9500061

root-mean-squared for random forest on training set

[1] 0.2452678

report R-squared on training data

[1] 0.7304121



Output 10: Geographically weighted regression

root-mean-squared for grw on training set

[1] 0.1625623

R-squared on training data

[1] 0.8837034

root-mean-squared for grw on test set

0.2329649

R-squared on training data

[1] 0.7697659

Test set proportion of variance accounted for by geographically-weighted regression = 0.770>

> # --------------------------------------

> # Construct a hybrid prediction

> # --------------------------------------

>

> houses.train$hybrid.pred <- (houses.train$rf.train.fit.full.pred +

+ houses.train$grw.train.fit.pred) / 2 # average of two best predictors

>

> houses.test$hybrid.pred <- (houses.test$rf.train.fit.full.pred +

+ houses.test$grw.train.fit.pred) / 2 # average of two best predictors

>

> cat("\n\nTraining set proportion of variance accounted",

+ " for by hybrid model = ",

+ sprintf("%1.3f",cor(houses.train$log\_value,houses.train$hybrid.pred)^2),sep=" ")

Training set proportion of variance accounted for by hybrid model = 0.935>

> cat("\n\nTest set proportion of variance accounted",

+ " for by hybrid model = ",

+ sprintf("%1.3f",cor(houses.test$log\_value,

+ houses.test$hybrid.pred)^2),sep=" ")

Test set proportion of variance accounted for by hybrid model = 0.813>