Report on expenditure prediction

Group Γ

April 30, 2013

Contents

1 Executive summary

Information on all cities in the state was available to predict expenditure per person for the forecasted data, based on the housing project developers' forecasts, for 2020 and 2040. Because of the way expenditure per person was distributed in the original data a log-scale transformation was used. A random forest analysis was used, wherein the computer uses algorithms to determine the best prediction. This leads to the predictions in the original scale as given in Table 1, which can be interpreted as the median optimal predictions. Of course there is uncertainty in the predictions and for that reason 80% confidence intervals for each estimate are also provided in Table 1. It is important to keep in mind that these predictions were based on the relationship in the past between Expenditure per person and Wealth per person, Population, Percent intergovernmental, Density, Income, and Growth rate. These relations may change between now and 2020 and 2040.

Table 1: Prediction and confidence intervals

City	Year	Point estimate	80 % confidence interval
Warwick	2020	254.0	168.0 - 475.7
	2040	260.0	168.0 - 489.2
Monroe	2020	241.0	159.0 - 368.8
	2040	238.0	162.0 - 355.6
Tuxedo	2020	464.7	275.0 - 795.0
	2040	444.2	290.1 - 637.8

2 Results

The point predictions and confidence intervals are given in Table 1. The predictions are the median values. Table 2 provides the current and predicted values for the 6 predictor variables and for Expenditure per person.

Table 2: Current and predicted city data

City	Year	Expenditure per person	Population	Wealth per person	% inter-governmental	Density	Income	Growth rate
Warwick	Now	237	16.225	78.908	24.7	170	19.044	30.3
	2020	254	20.442	85.000	24.7	214	19.500	35
	2040	260	31.011	89.000	26	325	20.000	40
Monroe	Now	159	9.338	55.067	8.8	599	16.726	30
	2020	241	10.496	58.000	8.8	695	17.100	35
	2040	238	13.913	60.000	10.1	959	18.000	35
Tuxedo	Now	926	2.328	155.034	6.1	52	30.610	2.5
	2020	465	10.685	116.000	6.1	249	28.300	300
	2040	444	29.246	115.000	7	656	25.000	100

Figure 1 demonstrates the results visually. The biggest change, with the biggest uncertainty, occurs in Tuxedo, where population size is forecasted to increase dramatically, and the Expenditure is forecast to decrease a lot before 2020. Note that the values on the Expenditure axes are not the same across plots, to increase the interpretability within city rather than across cities.

3 Methods and additional details

The comparison of the models was done with predicting expenditure per person on a natural logarithmic scale, i.e. we were predicting ln(expenditure). The reason for this is that some of the models assume normality of the residuals, and for expenditure there were a lot of small values and some large values. See Figure 2 for a histogram of expenditure and of ln(expenditure), which demonstrates the advantage of using the natural logarithmic scale. The cross validation coefficients are therefore an indication of the difference between predicted ln(expenditure) and actual ln(expenditure) for each datapoint. This means the final random forest analysis was also performed on the natural logarithmic scale, and the final results have been transformed back to the original scale by taking the exponent of the endpoints of the confidence interval and the predicted value. Because the endpoints as well as the point prediction, which is the median, are quantiles they can be transformed and still be interpreted as quantiles in the original scale.

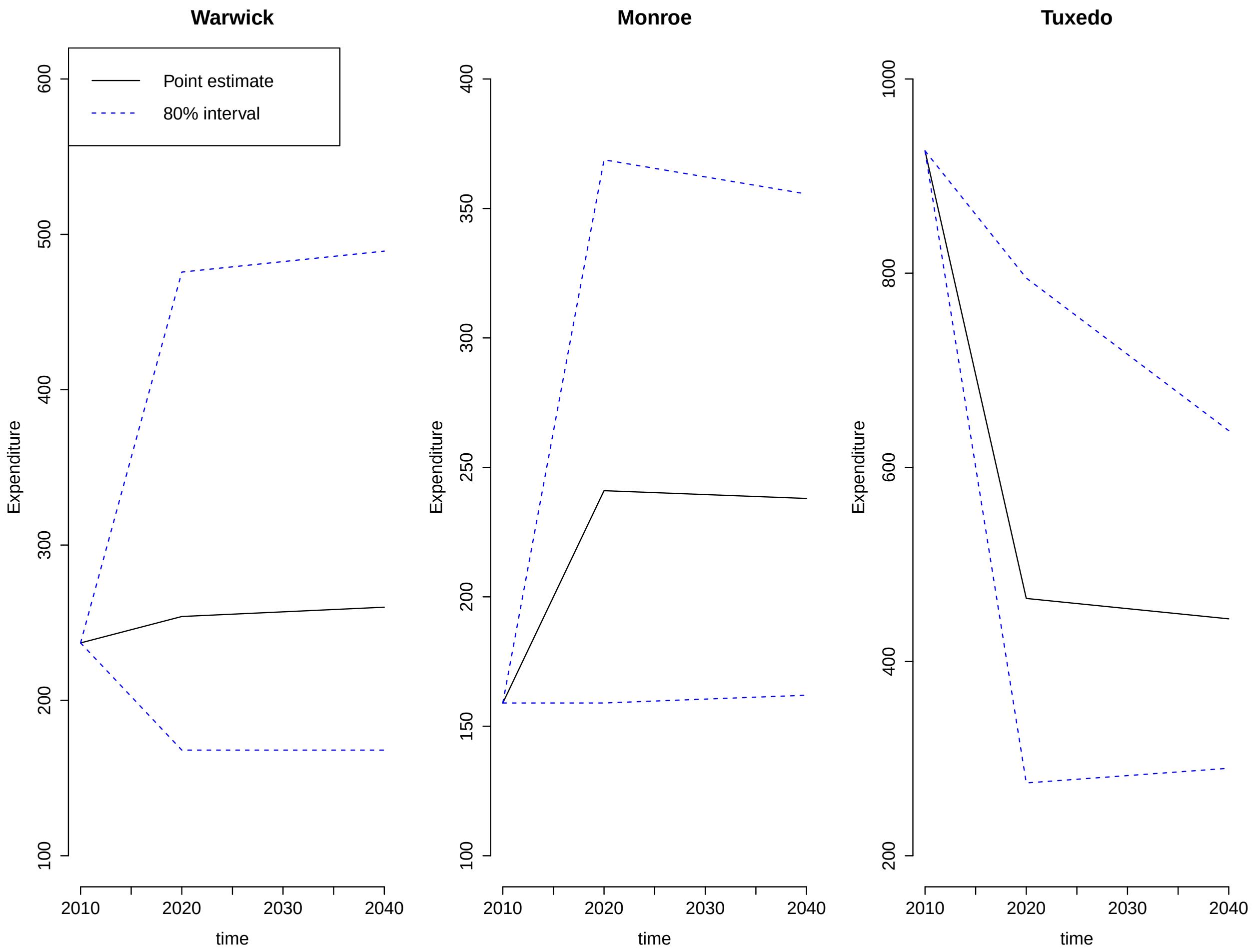

Figure 1: Point estimates (black) and 80% confidence intervals (blue) of Expenditure for each city

Several different models were compared in analyzing the data. It was decided that since our top priority is forecasting, we should use the model that exhibits the smallest amount of error in prediction. To compare the models, we used 11-fold crossvalidation. We split the datasets into 11 almost equally sized groups. Then we ran each model on 10 of the 11 groups, leaving one group out each time, once for each group. Then we used the model from the partial dataset to predict the log expenditure in the left out group. The "best performing" model was the model that let to the smallest squared differences (smallest sum of squared error) between the predicted value and known value (in log scale) for each left out group, summed across the eleven runs. The cross validation coefficients are given in Table 2. A lower coefficient is better, as it indicates a small difference between the predicted values and the actual values. The random forest analysis was chosen as the best predicting model.

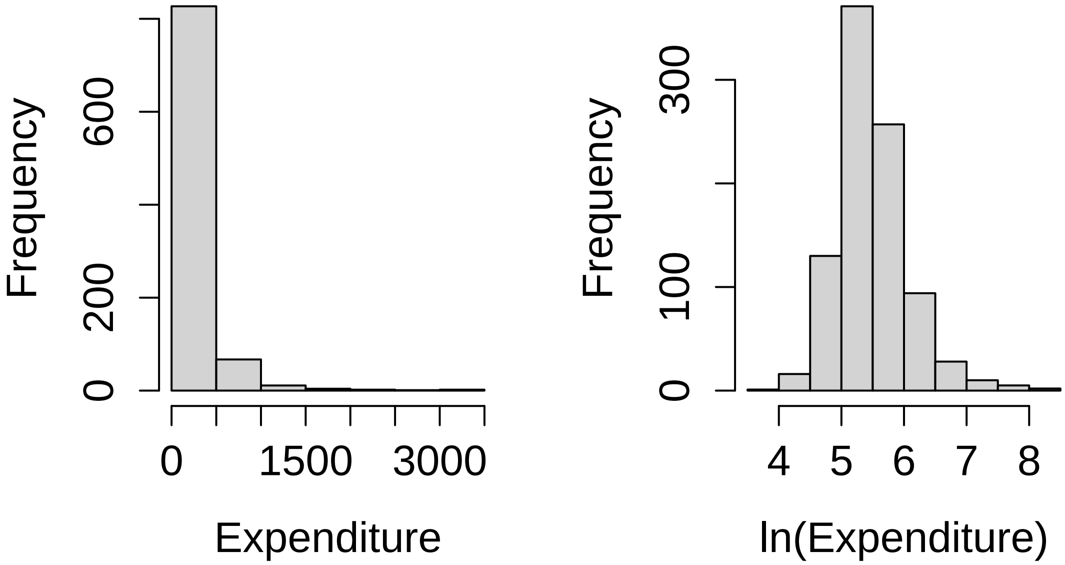

Figure 2: Frequency distributions of expenditure and ln(expenditure)

Random forest analysis is a machine learning method where the computer program creates a decision tree to use the predictors to predict the criterion. Such a decision tree can be started in different ways which results in different predictions. Random forest is called such because it creates a myriad of decision trees (a "forest"), starting in different random places. The point prediction is the median value predicted based on all these different trees, and the endpoints of the confidence intervals are the 10th and 90th percentiles of the values predicted by all these different trees.

Table 3: Model comparison

Model	Cross validation coefficients
Random forest	0.246
Local quadratic polynomial	1.416
Linear regression	1.362
Principal components regression	4.819