1-1-2016

Temperature Forecasts with Stable Accuracy in a Smart Home

Bruce Spencer
University of New Brunswick

Feras Al-Obeidat
Zayed University

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 726 – 733

Procedia Computer Science

The 6th International Conference on Sustainable Energy Information Technology (SEIT 2016)

Temperature Forecasts with Stable Accuracy in a Smart Home

Bruce Spencer[a,*], Feras Al-Obeidat[b]

[a]*University of New Brunswick, Fredericton, Canada*
[b]*Zayed Universty, Abu Dhabi, UAE*

Abstract

We forecast internal temperature in a home with sensors, modeled as a linear function of recent sensor values. When delivering forecasts as a service, two desirable properties are that forecasts have stable accuracy over a variety of forecast horizons – so service levels can be predicted – and that the forecasts rely on a modest amount of sensor history – so forecasting can be restarted soon after any data outage due to, for example, sensor failure. From a publicly available data set, we show that sensor values over the past one or two hours are sufficient to meet these demands. A standard machine learning method based on forward stepwise linear regression with cross validation gives forecasts whose out-of-sample errors increase slowly as the forecast horizon increases, and that are accurate to within one fifth of a degree C over three hours, and to within about one half degree C over six hours, based on one or two hours of history. Previous results from this data achieved errors within one degree C over three hours based on five days of history.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Temperature forecasts, forecast accuracy, domotic house, forward stepwise linear regression, service level agreement

1. Introduction

Machine learning, building on the shoulders of Artificial Intelligence and Statistical Inference, is capable of producing forecasts from information gathered from sensor networks. In this paper we consider forecast accuracy in a domotic or "smart" house, containing sensors. Such a house is described by Zamora-Martinez *et al.*[2] which competed in the Solar Decathalon 2012 competition[3]. It uses 88 sensors and 49 actuators, and records every 15 minutes the data collected from sensors. They also provided some of this data on the publicly available UCI data source[1]. That data consists of internal temperartures, lighting, CO_2 saturation, and humidity in various rooms, as well as external readings including temperature, humidity, wind speed, precipitation, atmospheric pressure, and sun on each external wall. This UCI data reports 21 of the 88 sensors, three of which have no data, leaving 18. The data is separated into two files, one covering four weeks of readings in March and April 2012 and the other covering two weeks of readings in April and May 2012.

* Corresponding author
E-mail address: bspencer@unb.ca

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2016.04.160

The homeowner can use accurate forecasts of internal temperature as part of an energy-management strategy to reduce energy consumption while maintaining internal temperature within a specified comfort range. For the SML house, described by Zamora-Martinez *et al.*[2] the Heating, Ventilation and Air Conditioning system (HVAC) system is responsible for 53.9% of the overall power consumption. As Pan *et al.*[13] state, due to thermal inertia, it is more efficient to maintain a temperature of a room or building than to heat or cool it. Therefore, considerable savings can be achieved by consulting future values of temperature and then deciding whether or not to activate the HVAC, rather than relying only on the present temperature. This was subsequently confirmed by the competition measures in the SML system prototype.[2]

Service-oriented architectures are the delivery method of choice for information technology businesses to provide value to clients. A natural candidate for a viable business model is to provide forecasts as a service from home sensor data. Like any business model, the client needs a guarantee of value for the cost of the service. In the case of forecasts, that guarantee is expressed as forecast accuracy. Statistical methods, such as those we describe in this paper, can help the service provider to establish guarantees with high probability. By monitoring the accuracy of the forecasts in a transparent way, the service provider and client can ensure the forecast quality meets service level agreements. When the client and server agree upon a forecast accuracy, a forecast horizon, and a subscription cost, the service can begin.

The client may need the forecasts to be tolerant of gaps in the series of observations due to outages in the sensor network. If the modelling technology requires an uninterrupted sequence of observations of a given length, say a few days[2], then that gap may result in no reliable forecasts being produced for that many days. If more outages occur while waiting, the waiting period restarts. Because the client is relying on forecasts to save energy, one important busniess consideration is whether a gap introduces a waiting period for forecasitng to restart, and for how long.

Using the setting and technology proposed here, once the service is set up, the client regularly provides the most recent sensor readings and receives forecasts covering the forecast horizon. The client sends readings from 18 sensors every 15 minutes and receives a forecast of the mean internal temperature in the living and dining room for the next six hours, *i.e.* 24 periods. We provide a method to measure forecast accuracy, validated by out-of-sample errors to within one fifth of a degree C over three hours, and within about one half degree C over six hours, In the case of any outages, the forecasts can restart after one or two hours, and in the mean time, previously delivered forecasts will be quite accurate for six hours.

In the remainder of this paper we investigate the use of stepwise forward linear resolution as a forecast methodology for this setting, which meets our requirements for accuracy and fast restarting, and therefore supports our service-oriented business model. Note that we do not advocate that this is the best modelling technology for this setting, only that it meets our minimal reqiuirements and is the only one we are aware of that does. The modelling method and results are presented and discussed in comparison to previous results with this data. We then summarize our results and propose future work.

2. Background: Temperarture Forecasting and Statistical Methods

According to recent studies, energy consumption in buildings represents 40% of the worldwide energy, of which more than a half is used by HVAC systems.[9,10] Accurate temperature forecasts can reduce these costs. Model Predictive Controllers (MPC) are characterized by the explicit use of a process model in order to obtain the control signal while minimizing a cost function. The cost function takes into account a prediction horizon of a given length and a control horizon[7]. Álvareze *et al.*[7] describe a house that is fed only by renewable sources of electricity, and is equipped with a variety of sensors, including temperature, CO_2 saturation and humidity in various rooms, and a fancoil actuator which is a part of the HVAC system. It employs a Practical Non-Linear Model Predictive Controler (PNMPC) which includes an optimizer, a model of the fancoil, a model to simulate the thermal conditions of the room, and an index to evaluate the thermal comfort of the household users. The objective is to control the thermal conditions in order to maintain the thermal comfort of the users. The systems produces good behavior and has been able to maintain the thermal comfort.

We analyse data provided by Zamora-Martinez *et al.*[2], which reports 18 sensors, to create forecasts of the internal temperature. However, they use only two of the sensors: internal temperature and sun irradiance. They also encoded the hour as 24 Booleans. The data is presented as a time series to an online learning framework. From an initially uninformed model, each new set of sensor readings is loaded and the model is improved. After 5 days of data, about

480 observations, the forecasts show good accuracy. The modelling technology uses an artificial neural network, which is trained using a variety of mechanisms that are variants of gradient descent and Bayesian linear models. Accuracy for 12 forecasts, *i.e.* 3 hours, is calculated for each of the methods over a number of experiments. Bayesian linear performs the best most often, and the forecasts from this technique have MAE below 0.2° C about half of the time, while its error can be up to 1°C.

In a service-oriented setting, the client provides historical sensor data to a service and receives from the service a table organized by forecast horizon showing the projected forecast errors and costs to the client. The service computes this table from the data, by selecting a modelling method with associated hyperparameters. Using statistical techniques, the service estimates the forecast errors and computing costs. The forecasting method we use is a variant of linear resolution known as forward stepwise linear resolution, as presented by Hastie *et al.*[4], and provided the `leaps` package, using the R method `regsubsets`. As with any linear regression problem, we are given a set of independent variables $x_1, ..., x_n$ and a dependent variable y of interest that we want to forecast as a function of the independent variables. That is, we seek parameters $\beta_0, ..., \beta_n$ so that $\beta_0 + \Sigma_{i=1}^{n} \beta_i x_i$ is a good approximation of y. When presented with a set of m instances of each x_i, called $x_{i,j}$ and the corresponding instances y_j, we select the β_i parameters so that the residual sum of squares $\Sigma_{j=1}^{m} (\beta_0 + \Sigma_{i=1}^{n} \beta_i x_{i,j} - y_j)^2$ is minimized. Stepwise forward regression initially sets β_0 to $\bar{y}$, and all other $\beta = 0$. Then it repeatedly selects i and a value for β_i so that the error function is reduced as much as possible. Once a value of β_i is selected, it is not changed further. After all such β_i are selected, stepwise regression halts with the model.

Minimizing a model's error is not the only criteria for choosing a model. If it were, then a model could be created with zero error on the given data, simply by replicating the y for each vector x. But such a model would have one parameter for each m, *i.e.* it would be as large as the input data. It would likely perform poorly on out-of-sample data that were not available when the model was created, *i.e.* it would be overfit. Even if the model size is n, the number of independent variables, which is often smaller than m, the model may be overfit. The measure of a model's error that we use is the Bayesian Information Criteria (BIC)[8], defined as $1/n$ (RSS + $log(n)d\hat{\sigma}^2$) where n is the number of observations, d is the number of dimensions of the model which in our cases is the number of non-zero β's, and $\hat{\sigma}^2$ is an estimate of the variance of the variable of interest. BIC penalizes larger models, and thus balances model size against forecast error.

3. Experiments and Results

We experiment with both one and two hours of historical data, and for each we generate forecasts for each of the 24 periods, *i.e.* the next six hours. Specifically, we build a model to predict the internal temperature at some time steps f into the future, based on looking at the current data at each of 18 sensors, and looking into the past b time steps for each sensor. That is, one row of our data table contains one internal temperature f forward time steps into the future, and 18 sensor observations taken at $0, ..., b$ times steps into the past, for a total of $1 + 18(b + 1)$ variables. Those predictor variables recalled from previous times are called *lagged* observations. We will build one model for each f time steps into the future, so f will range from 1 to 24. Let the lag l vary across the time steps 0 through b into the past, and let k vary across the 18 sensors. Let $x_{k,t}$ be the t^{th} observation for the sensor k counting from the first observation at $t = 1$. Let y_t be the mean internal temperature of the house at time t. For each f we seek to choose the $1 + 18(b + 1)$ values for the intercept $\beta_{f,0}$ and each coefficient $\beta_{f,k,t}$ for the value of the k^{th} sensor at time t. We want to minimize the error function

$$\Sigma_{t=b+1}^{m} (\beta_{f,0} + \Sigma_{l=0}^{b} \Sigma_{k=1}^{18} \beta_{f,k,t-l} x_{k,t-l} - y_{f+t})^2 \quad (1)$$

In this equation, t starts at b+1 because there are not enough observations to provide values for the lagged readings for the first b data points. We perform two tests, selecting b as either 4 and 8, to use either one or two hours of observations, so there are either 5×18 or 9×18 predictor variables. The model for a given f and b consists of the $1 + 18(b + 1)$ values for β. It is built using 2/3 of the data available, and then validated using the remaining1/3 of the data. Because we use stepwise forward regression, initially $\beta_{f,0}$ is the mean of the observed internal temperatures $1/(m - b)\Sigma_{j=b+1}^{m} y_{f+j}$. All other β are initially zero, and they are estimated one by one until all are estimated, as described in Section 2. While each step of the foreward regression improves the error over the previous step, the amount of improvement decreases. Many of these computed β could just as well be left at zero, since the model will become affected by less and less significant phenomena and not perform well in different circumstances; *i.e.* it will have a large number d of dimentions and will be overfit. For each regression step, we measure the model using the BIC, which considers the complexity of the model as well as its error. We record the BIC for the model of each size, where size is number of steps that the foreward regression has made. Since we are using 10-fold cross validation, we run the stepwise regression once for each fold. This gives us 10 estimates of the BIC for each model size. We consider the mean of these estimates as a good predictor of the BIC for a given model size. Since BIC reduces and then increases, we can identify the model with minimal BIC. But since the BIC decreases slowly as it approaches the minimal, we are interested the model size whose mean BIC is within one standard error of smallest of this minimal model. See Figure 1, where the estimated BIC for each model size is shown, and marked with a confidence interval spanning one standard error, based on the BIC's of the models of that size over the 10 folds. The model size with the smallest mean BIC is shown labeled green, and the model whose BIC is larger by at most one standard error is labelled red. The size of this red model is recorded as the appropriate size for a regression model for this data. We show the model's BIC measures for $f = 1, ..., 6$ hours forward and $b = 8$ for two hours back. Other diagrams for various b and f are similar. The model sizes vary from 10 to 24 for the first 20 intervals, which is a good reduction since the maximal model size is 163. For the later forecasts, Specifically for Data Set 1 the model sizes for each of the 24 forecasts are 10, 14, 16, 16, 16, 20, 18, 19, 18, 20, 20, 19, 21, 23, 20, 20, 22, 20, 20, 19, 24, 23, 23, 24.

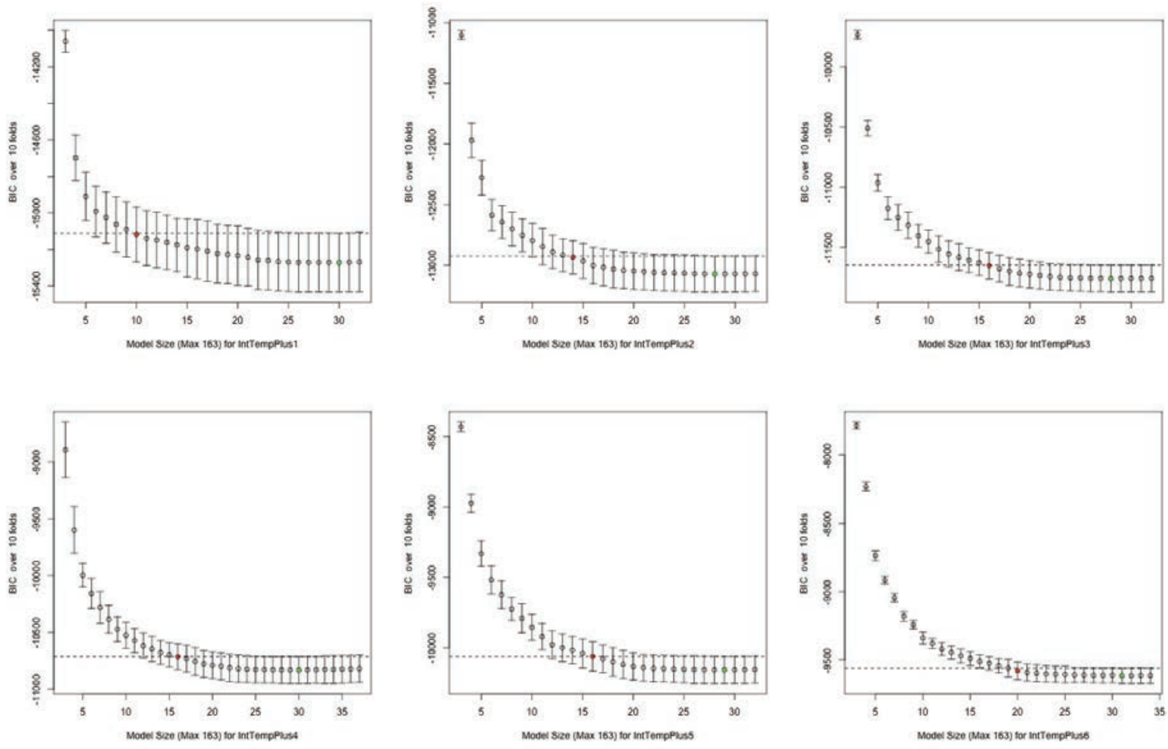

Fig. 1: Selecting the number of predictors based on BIC for first 6 models using Data Set 1. Others figures are similar for both Data Sets. These figures show that the model sizes are 10, 14, 16, 16, 16 and 20 predictors, reduced from a maximally possible 163.

Shown in Table 1 is the set of predictor variables for which β is non-zero. Ignoring the β for the intercept, in the first row, there are 10 predictor variables selected to forecast the internal temperature one period ahead. Here *TD* is the temperature of the dining room, *TL* is the temperatire of the living room, *TW* is the weather temperature, possibly from the weather forecast (the data description is not explicit about its meaning), *CD* is the CO_2 staturation of the dining room, *CL* is the CO_2 staturation of the living room room, *HD* is the humidty of the dining room, *HL* is the humidity of the living room, *LD* is the lighting of the dining room, *LL* is the lighting of the living room, *Pcp* is the

Table 1: Models for DataSet 1 with Lag 4

Int	*TD*	*TL*	*TW*	*CD*	*CL*	*HD*	*HL*	*LD*	*LL*	*Pcp*	*Tw*	*Wi*	*SW*	*SE*	*SS*	*P*	*T*	*H*	RMSE	MAE
1	0, 1, 2	0, 1, 3							0					2			0, 1		0.048	0.019
2	0, 1, 3	0, 1, 4		0, 1	0									4		0, 4	0, 1		0.053	0.038
3	0, 1, 2	0, 1, 4		0	0				0				1	0, 3		0	0, 4		0.186	0.059
4	0, 1	0, 1, 4		0, 1	0, 2	4		0, 4		0			0	0, 4		0	0, 4		0.776	0.085
5	0, 1	0, 1, 4		0, 1	0	0		4		0	0		0, 4	0	0		0, 4	1	1.812	0.108
6	0, 1	0, 1, 4			0, 2	4		4	0	0	0		4	0	0		0, 4	4	2.108	0.134
7	0, 1	0, 3, 4			0	0		4		0	0		1, 4	0, 2	0		0, 1, 4	2	2.887	0.168
8	0, 1	0, 4		0, 1	0, 3	3		4	2	0	0		3	0, 1	0		0, 4	3	3.665	0.2
9	0, 1	0, 3, 4			0	0		4	2	0	0		1, 4	0	0, 2		0, 4	0	4.73	0.24
10	0, 1	0, 4				0	0		0	0	0, 2		2, 4	0			0, 1, 4	0	3.904	0.275
11	0, 1	0, 4				0			0	0	0, 2		2, 4	0			0, 1		5.208	0.338
12	0, 1	0, 4				0			0	0	0		0, 4	0			0, 1		6.291	0.392
13	0, 1	0, 4				0			0	0	0, 1		4	0			0, 1		7.466	0.446
14	0, 1	0, 2, 4				0			0	0	0, 1		4	0			0, 1	1	8.505	0.494
15	0, 1	0, 2, 4				0				0	0, 1		0, 4	0, 1			0, 2		9.979	0.558
16	0	0, 1, 4				0				0	0, 2		0, 4	0			0, 2		11.803	0.631
17	0, 1	0, 1, 4				0				0	0, 3		0, 4	0			0, 2		12.057	0.657
18	0, 4	0, 4	4			0				0	0, 2		4	0		4	0, 2		13.64	0.743
19	0, 4	0, 2	4			3		3		0	0, 3		4	0			0, 2, 4	2	15.239	0.821
20	0, 1	0, 4	4			3		2		0	0, 3		4	0			0, 2	2	17.058	0.894
21	0, 4	0, 4	4			4				0	0, 1, 3		2, 4	0	0			0	22.298	1.038
22	0, 4	0, 4	4		0	4				0	0, 1, 4		2, 4	0	0	4		0	23.036	1.094
23	0, 4	0, 4				4				0	0, 3		4	0	0			0	25.242	1.18
24	0, 4	0, 4	4			4				0	0, 3		4	0	0			0	27.082	1.256

Table 2: Models for DataSet 1 with Lag 8

Int	*TD*	*TL*	*TW*	*CD*	*CL*	*HD*	*HL*	*LD*	*LL*	*Pcp*	*Tw*	*Wi*	*SW*	*SE*	*SS*	*P*	*T*	*H*	RMSE	MAE
1	0, 1, 3	0, 1, 6							0					2			0		0.056	0.019
2	0, 1, 2	0, 1, 8		0	0			6					1	0, 4		0			0.153	0.042
3	0, 1, 2	0, 1, 8		0	0			6					1	0, 3		0	0, 5		0.028	0.061
4	0, 1, 2	0, 1, 8							0		0		0	0, 4		0, 6	0, 6		0.277	0.077
5	0, 1	0, 1, 6, 8			0			4	0		0			0	0	6	0, 6		1.009	0.096
6	0, 1	0, 1, 6, 8			0	2		4		0	0		1	0	0	7	0, 2, 8	0	2.338	0.135
7	0, 1	0, 3, 8				3		4		0	0		1	0, 2	0	6	0, 8	1	3.244	0.172
8	0, 1	0, 1, 4, 7				5		5		0	0		3	0, 1	0	8	0, 4	2	3.07	0.195
9	0, 1, 7	0, 3, 8			0	3				0	0		1	0	0	7	0, 1	0	3.669	0.234
10	0, 1, 8	0, 4, 8			0	3	0		0	0	0		2	0	0	6, 8	0, 1		4.246	0.275
11	0, 1	0, 5, 8				0			0	0	0, 3		1, 5, 8	0	8	0, 8	0, 1		6.253	0.351
12	0, 1, 8	0, 5				3			0	0	0, 3		1, 4, 7	0	7	7	0, 1		7.36	0.409
13	0, 1, 8	0, 6			0	5			0	0	0, 3		1, 4, 7	0	0	7	0, 1, 8		7.587	0.438
14	0, 8	0, 2, 8		3		0			0, 8	0, 8	0, 3		4, 8	0		6, 8	0, 1, 8	1	7.141	0.489
15	0, 8	0, 1, 6	8			0			8	0, 8	0, 3		4, 8	0	8	8	0, 2		12.148	0.619
16	0, 8	0, 1, 8	8			0			7	0, 8	0, 3		4, 7	0		7	0, 2, 8		10.905	0.637
17	0, 8	0, 1, 8	8		1	0			6	0, 8	0, 1		4, 6	0, 1		6	0, 2, 8		10.866	0.666
18	0, 8	0, 8	8			0		8		0, 8	0, 1		4, 6	0, 1		6	0, 2, 8		13.762	0.755
19	0, 8	0, 8	8			7			7	0	0, 1, 8		4, 8	0		5, 7	0, 8	8	15.241	0.826
20	0, 8	0, 8	8			7				0	0, 1, 8		4, 6	0, 1		6	0, 8	7	15.059	0.878
21	0, 8	0, 3	8		0	5			6, 8	0, 8	0, 3		4, 8	0, 4	0	6, 8	0, 8	0	17.787	0.979
22	0, 8	0, 3	8		0	5			6	0, 8	0, 3	0	4, 8	0, 4	0	6	0, 8	0	18.963	1.034
23	0, 8	0, 3	8	8	0	7	5		7	0, 8	0, 3, 8	0	5, 7	0	0	7		0	24.485	1.18
24	0, 6	0, 3	8		0	7			1	0, 8	0, 3, 8		5, 8	0	0	4, 8	0, 8	0, 8	23.619	1.179

amount of precipitation, *Tw* is the twilight indicator, *Wi* is the external wind speed, *SW* is the amount of sun on the west external wall, *SE* is the sun on the east external wall, *SS* is the sun on the south external wall, *P* is the sun's irradiance measured on a pyranometer, *T* is the external temperature and *H* is the external humidity. In the first line, we can see that current dining room temperature and the dining room temperature 15 minutes ago, the current living room temperature, the living room temperature 15 minutes ago and 30 minutes ago and other sensors at other lags, are all selected as contributing to what the temperature will be in 15 minutes from now. The recent sun on the east wall, the recent lighting in the living room and the recent dining and living room CO_2 saturations are also deemed useful. In fact the regression formula with the variables ordered by error reduction is

$$Int_1 \sim (Intercept), TL_0, TD_0, TD_1, TL_1, SE_2, LL_0, TL_6, TD_3, T_0.$$

Table 3: Forecast Errors for each Data Set and each Lag

	(a) Data Set 1 Lag 4		(b) Data Set 1 Lag 8		(c) Data Set 2 Lag 4		(d) Data Set 2 Lag 8	
Horizon	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1	0.0476	0.0189	0.0562	0.0192	0.1437	0.0134	0.061	0.0119
2	0.0502	0.0285	0.1151	0.0306	0.2382	0.0201	0.2699	0.0201
3	0.1147	0.0387	0.0954	0.0407	0.347	0.0275	0.398	0.0281
4	0.4002	0.0503	0.1614	0.0499	0.5169	0.0365	0.6432	0.0384
5	0.8856	0.0619	0.4737	0.059	0.737	0.0465	0.7405	0.0467
6	1.1803	0.0738	1.0475	0.0716	0.9969	0.0579	0.9275	0.0573
7	1.5438	0.0873	1.5631	0.086	1.6984	0.078	1.1677	0.0686
8	1.9399	0.1014	1.8206	0.0996	2.0472	0.0934	1.5273	0.0837
9	2.4139	0.1167	2.1072	0.1145	2.9757	0.1204	1.901	0.0988
10	2.6013	0.1326	2.4078	0.1305	3.6273	0.1442	2.3641	0.116
11	2.935	0.1513	2.9695	0.1505	4.2809	0.1689	2.952	0.1362
12	3.345	0.1713	3.5482	0.172	4.9121	0.1943	3.2723	0.154
13	3.8221	0.1924	4.0054	0.1924	5.3788	0.2169	3.8039	0.1749
14	4.3271	0.2138	4.305	0.2135	5.8794	0.2401	4.4258	0.1983
15	4.909	0.2367	5.2072	0.2405	6.486	0.2653	4.9528	0.2211
16	5.5928	0.2613	5.7305	0.2652	7.1227	0.2923	5.5761	0.247
17	6.1622	0.2845	6.151	0.2887	8.1132	0.3258	6.3116	0.2752
18	6.795	0.3098	6.7992	0.3145	8.9972	0.3589	7.0736	0.3046
19	7.4788	0.3366	7.4829	0.3414	9.7231	0.3909	8.1295	0.3408
20	8.2248	0.3644	8.0312	0.3681	10.541	0.4246	9.1282	0.3768
21	9.3819	0.3963	8.7436	0.3971	11.0574	0.4541	9.9817	0.4124
22	10.3954	0.4279	9.4486	0.426	12.1884	0.4945	10.9673	0.45
23	11.4448	0.4605	10.5529	0.4586	13.4012	0.5381	11.8862	0.4878
24	12.4888	0.4935	11.3969	0.4885	14.9303	0.587	13.1502	0.5316

Regressing on these 9 variables over the 2/3 training data generates a formula with a mean absolute error of 0.019° C, and a mean squared error of 0.056. These are errors in the training set.

Looking over Tables 1 and 2, we can see patterns emerging. Certainly the temperatures of the living and dining rooms are useful predictors of the mean internal temperature, which is in fact just the mean of the living and dining room temperatures. The humidity of the dining room is more important than the humidity of the living room. The external wind is not important but the precipitation is, especially the precipitation about one hour ago. The sun on each wall is important. It is interesting to note how little benefit for both MAE and RMSE arises when considering the sensor values more than one hour ago. The MSE and MAE are not significantly lower in the Table 2 than they are in Table 1.

Table 3, shows the forecast errors over the validation set for the remaining 1/3 of the data, and a forecast from the first period up to the given horizon, both as MAE and RMSE. In the forecast as a sevice setting, we use these observed forecast error to predict what service the client could expect to achieve, and this table can be used to predict expected forecast error. For instance, if the client had given us Data Set 1, and needs the MAE to be below 0.15° C, he would not be able to use a forecast horizon longer than 2.5 hours, according to this table. We also show the results for the second data set, which are similar.

4. Discussion

Our intuition led us to surmise there would be advantages to using many sensors to capture the effects of people's activities in the house. A comparison with previous results on this data bears some evidence that this intuition is correct, but let us consider it further. We can classify the influencers on internal temperature as time-dependent, such as solar irradiance, and time-independent such as lighting and CO_2 stauration which are associated with random human activities such as people occupying rooms and performing activities such as cooking. To forecast properly requires taking both into account.

There are a variety of ways to include both time-dependent and time-independent data in one model, including multivatiate time series analysis[5] which analyses the regular (seasonal) patterns in the time series and correlations between the variables. Another is by using lagged variables, *i.e.* replicating any historical sequence of values which has been offset or *lagged* by a number of time periods, and including this offset sequence as a new predictor variable in the model. (Although commonly used, it appears to have been first done with regression by Durbin[6].) This paper offers an initial investigation of this second option where we choose all 18 predictor variables for replication and we make either 4 or 8 replications of each. Since we introduce 4×18 or 8×18 new predictor variables in the model, we need to eliminate those predictor varuables that do not contribute very much to the forecast, in order to avoid over-fitting, and this paper proposes a simple way to identify those variables that are not time dependent. However, there may be more reliable ways.

Our intuition also led us to surmise that the effect of time-dependent activities on internal temperature might be modelled accurately by looking back only one or two hours, rather than a large number of periods. Since temperature is governed by the second law of thermodynamics operating in a small space, influencers on temperature will quickly have an effect, and these effects will have a short duration as more recent influencers will replace them. If we were to know all of the temperatures of all of the objects in a room in one instant, in theory we could compute a near exact room temperature in the next instant, removing the need for any historical values of variables. Even though we are not given this much data, we use this principle and choose a small amount of historical data. Relying on immediate or almost immediate observations alone is not enough. We can and should exploit time dependent factors, *i.e.* factors that occur according a schedule. For instance, the effect of sunrise can be predicted and is dependent on the hour of the day, with a seasonal variation. In this particular case, we actually do not need to rely on the scheduled time of day that dawn will occur. We instead have access to a light sensor on the external east wall that will herald an increase in temperature due to sunrise before its effect occurs inside the house. In other words, a light event in the east will preceed a heating event, so we can forecast a heating effect of the post-dawn sunlight without needing to know the time that dawn occurs, as long as we are made aware of it by the pre-dawn light. So we use the lagged light signal, *i.e.* the light from the east from several time steps ago to get the same effect as the solar pattern. While one cannot be guaranteed that all time-dependent events will be associated with both a schedule and a precursor, in this investigation we limited ourselves to looking at the current values and the lagged values of sensor data for all sensors in the network. We therefore use no time-dependent factor. This essentially prevents us from noticing time-dependent patterns. Where this simple technique will fail is predicting something that occurs on a schedule with no sensor that heralds that event. For instance, if a timer is set to turn on a heater in the home every day at a fixed time and if no sensor captures this trigger, then our method would miss the infliuence of this heater on the internal temperature. Thus in future modelling scenarios, it may be important to include the time signal.

Our forecast errors validate our two intuitions. Our model uses all 18 available sensors and replicates them 4 or 8 times, in separate experiments. In both experiments our errors are within about $0.17°$ C for a three hour horizon, and within about $0.5°$ for a six hour horizon for Data Set 1. On the other hand, Zamora-Martinez *et al.*[2] provide not 18 but instead two sensors, internal temperature and sun irradiance, and about 5 days of history. There forecasts are within about $1°$ C over a three hour horizon.

Linear regression is not very much affected by missing observations. In a real-world setting, it is quite likely that a sensor will malfunction and its readings will be unavailable quite regularly. Given the most recent reliable regression model based on recent data, where the regression equation uses lagged observations over, say, two hours, our system would be able to resume producing forecasts two hours after the sensor was repaired, and meanwhile could rely on previous forecasts for 6 hours. On-line learning systems require a long training period, and when faced with a long gap in the data from a sensor it relies upon, it would have to start again from an initial state of zero knowledge. Thererfore it would require a long delay to start producing reliable forecasts, which could be up to 5 days, or longer if another outage were to occur, and during most of this time, no forecast would be available from this source.

To calculate Table 1 on a 2.5 GHz Mac running OSX 10.11.3 with 16 GB of memory takes about 12 CPU seconds, while Table 2 takes about 100 CPU seconds.

5. Conclusion and Future Work

Our goal is to generate forecasts of the internal temperature in a home with access to data from a variety of sensors. Our main success criterion are forecast accuracy, and the ability to restart forecasts quickly after a sensor failure so that energy savings are not delayed. A sequence of regression steps is guided by a greedy algorithm: forward stepwise linear regression. Model size is limited by estimating its BIC over 10-fold cross validation. The BIC metric balances smallness against accuracy. We conservatively choose the minimal sized model with BIC within one standard error of the minimal BIC estimate. High out-of-sample accuracy over a holdout set of 1/3 of the data indicates we meet our main criterion. Forecast accuracy is also affected by the number of historical readings used in each forecast, although 8 historical readings are not much better than 4. Therefore, forecasting can restart 75 minutes after an outage, achieving our second criterion. Accuracy is observed to reduce gradually as the forecast horizon increases, so one can tradeoff better accuracy against a longer horizon. This supports service contract negotiation between the forecast server and the forecast consumer.

Our simple modelling approach demonstrates that the sensor data can support a valuable service for the customer. In future work, we will consider increasing this value. When adding more historical data, the accuracy increases only marginally, which suggests that this problem is not a good candidate for time series analysis. However this theory needs to be tested. Using a time series may might offer some benefit by modelling all future temperatures with one model, whereas now we forecast each future temperature separately. We will investigate variations of multivariate time series, ARIMA models, other types of artificial neural networks [12], autoregressive modelling with exogenous variables (ARX) [11] on this same data. We will also investigate of how time-dependent and time-independent variables can be distinguished and how each should be treated. We currently select features by balancing model size against accuracy gains in a greedy algorithm. We will experiment with other methods to extract features and to selectively define new features.

References

1. https://archive.ics.uci.edu/ml/datasets/SML2010.
2. F. Zamora-Martínez, P. Romeu, P. Botella-Rocamora, J. Pardo, *On-line learning of indoor temperature forecasting models towards energy efficiency*, Energy and Buildings, Volume 83, November 2014, p. 162-172, ISSN 0378-7788,
3. United States Department of Energy, Solar Decathlon Europe Competition, 2012, http://www.solardecathlon.gov.
4. T. Hastie, R Tibshirabi, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction, Second Edition*, Springer, 2011, ISBN 387848576.
5. R. Tsay, *Multivariate Time Series Analysis: With R and Financial Applications*, Wiley, 2013, ISBN: 978-1-118-61790-8.
6. J. Durbin, *Estimation of Parameters in Time-Series Regression Models*, Journal of the Royal Statistical Society, Series B, (Methodologicla) Vol. 22, No. 1, 1960, p. 139-153, http://www.jstor.org/stable/2983884.
7. J.D. Álvarez, J.L. Redondo, E. Camponogara, J. Normey-Rico, M. Berenguel, P.M. Ortigosa, *Optimizing building comfort temperature regulation via model predictive control*, Energy and Buildings, 57, (2013) 361 - 372.
8. A. A. Neath, J. E. Cavanaugh,*The Bayesian information criterion: background, derivation, and applications* Wiley Interdisciplinary Reviews: Computational Statistics, Volume 4, Issue 2, pages 199203, March/April 2012.
9. P.D. Morosan, R. Bourdais, D. Dumur, J. Buisson, *Building temperature regulation using a distributed model predictive control*, Energy and Buildings 42 (2010) 14451452.
10. L. Pérez-Lombard, J. Ortiz, C. Pout, *A review on buildings energy consumption information*, Energy and Buildings 40 (2008) 394398.
11. A. Mechaqrane, and M. Zouak, *A comparison of linear and neural network ARX models applied to a prediction of the indoor temperature of a building* Neural Computing & Applications, 13(1), 3237, (2004).
12. B. Thomas, and M. Soleimani-Mohseni, *Artificial neural network models for indoor temperature prediction: Investigations in two buildings* Neural Computing & Applications, 16, 8189 (2007).
13. D. Pan, Y. Yuan, D. Wang, X. Xu, Y. Peng, X. Peng, P.-J. Wan, *Thermal inertia:towards an energy conservation room management system*, in: A.G. Greenberg,K. Sohraby (Eds.), INFOCOM, IEEE, 2012, pp. 2606 - 2610.