Energy Forecasting for the Global Energy Forecasting Competition 2014

Semester Project Report

Fabian Brix
MSc Candidate
School of Computer & Communication Sciences
Swiss Federal Institute of Technology Lausanne
fabian.brix@epfl.ch

Abstract—The abstract goes here.

I. INTRODUCTION

A. GEFCom 2014

The Global Energy Forecasting Competition (GEFCom 2014) is the second edition of a competition first held on Kaggle in 2012 that attracted hundreds of participants contributing many novel ideas to the energy forecasting field. The second edition lasted from 08/15/2014 to 12/15/2014 and was sponsored by several IEEE bodies and the International Journal of Forecasting. It included four competition tracks: electric load, electricity price, wind power and solar power forecasting. The respective data was published on the community platform of crowdanalytix.com on a weekly basis during the competition. The format of the competition was that of rolling forecasting requiring contestants to submit the next period of interest to forecast a day before the next set of data was published. To appear on the final leaderboard the contestants had to submit 99 quantiles for each step throughout the forecast horizon. In the course of this semester project we focused solely on the electric load forecasting track. This track had a weekly forecast horizon of one month and consisted of three trial periods and twelve competitive periods. In order to feature on the final leaderboard nine competitive submissions were required. Due to our lack of familiarity with the subject, participating in the competition was not prioritized after initial attempts. Instead the focus was on comparison of the performance of different methods on the whole dataset via cross validation.

B. Short Introduction to Energy Load Forecasting

Before we review related work we are going to introduce the area of Energy Load Forecasting. Load forecasting, as it is commonly referred to, is usually concerned with the prediction of hourly, daily, weekly, and annual values of the system demand and peak demand of an electric utility [**?**]. Such forecasts are sometimes categorized as short-term (up to 1 week), medium-term (1 week - 1 year) and long-term ($¿$ 1 year) forecasts, depending on the time horizon. In the load forecasting track of GEFCom 2014 we are concerned with forecasting the daily electricity load of a utility for a whole month on a rolling basis given the data of the previous years. The task is therefore on the threshold between short-term and medium-term load forecasting.

Load Forecasting is necessary for the planning of energy system and their effective operation and maintenance. Forecasting accuracy therefore has a major impact on electric utilities and their regulators. In case of overestimation of future energy load, utility providers will operate too many units possibly driving energy demand and in case of long-term forecasts investment in the construction of new infrastructure can be wasted. Underestimation leads ot unmet demand and systems that are vulnerable to crashes. The output of such forecasts can either be point forecasts or estimates of the probability distribution of values of future demand or load as required during GEFCom 2014. Electricity load follows a nonlinear, volatile pattern subject to several exogenous variables such weather conditions, randomness in human behavior leading to randomness in demand and economic conditions and demographic changes. In GEFCom 2014 the exogenous variables are limited to weather conditions in the form of recorded temperature at several sites and calendar effects such as the effects of weekends and holidays on the electricity demand. In this report we analyse the predictions produced by algorithms that are capable of capturing the nonlinear dependencies between these exogeneous variables and the load such as General Additive Models, Random Forests, Multilayer Perceptrons. We do not strive to build any algorithms ourselves, but employ the various implementations in R of the mentioned algorithms.

LOAD vs DEMAND?? == consumption vs. demand??

II. REVIEW OF RELATED WORK

Get relevant sections from Tao Hong's working paper literature review Day of the Week etc. Hyndman for more features Review previous work from Gefcom 2012 and beyond including papers with related approaches including but not necessarily limited to those that can be found under the following link: http://blog.drhongtao.com/2014/08/recommended-papers-for-gefcom2014-contestants.html. Distinguish between approaches to temperature and load prediction.

Probabilistic Electric Load Forecasting: A Tutorial Review contains literature review for longterm probabilistic load forecasting

Tao Hang, Jason Wilson: Long Term Probabilistic Load Forecasting and Normalization With Hourly Information but papers is on Long Term Electric Load Forecasting: 1-X years [1]

Rob Hyndman, Shu Fan Density Forecasting for Longterm Electricity Demand Source for calendar effects [2]

Shu Fan, Rob Hyndman Short-term load forecasting based on a semi-parametric additive model [**?**]

III. Dataset and Evaluation Metrics

A. Dataset

The dataset provided by GEFCom 2014 includes hourly historical load and weather data of a utility in an undisclosed area on the east coast of the United States of America. The 25 weather stations in the dataset provide historical temperature for their respective zones. However, the load data consists only of the system level load in Mega Watts (MW) and not all of the zonal level load series. Therefore, forecasts in the context of Smart Grid Technology are not required. The temperature data made available consists of 25 series of temperature data in Fahrenheit from the 25 different weather statios dating from 01/01/2001 to 12/01/2011. The load data of the utility is recorded starting from the 01/01/2005 at 1am.

The complete data set acted both as training and validation set except for the first and last month of the data published for the reason that the data was released on a weekly basis. It consists of 15 spreadhseets in the format of Comma-Seperated Values (CSV) for the load forecasting track. The first spreadsheet contains data starting from 01/01/2001 up until midnight on the 10/01/2014 at 1am from when on the incremental spreadsheets released every week contain only one month of data.

The forecasts were required to be made starting from 10/01/2010 on a monthly rolling basis for 15 months. The nature of the data provided required the contestants to produce their own temperature forecasts for the month ahead in the dataset.

LOAD: one utility! Domestic + Industrial –¿ less nice patterns? –¿ less impact of holiday effect?

B. Evaluation Metrics

The comptetition requries each the participants to provide percentile forecasts, that is forecasts of the quantiles $\tau = 0.01, 0.02, \ldots, 0.99$ with natural lower and upper bounds. The Evaluation Metric employed to score the contestants' submissionsis the tilted loss/error function also known as the pinball loss/error function. In the following paragraphs let y denote an observation and $\hat{y}$ denote a corresponding forecast while ξ is defined as the residual $y - \hat{y}$.

$$L_\tau(\xi) = \begin{cases} \tau\xi & \text{if } \xi \geq 0 \\ (\tau - 1)\xi & \text{if } \xi < 0 \end{cases} \quad \text{where } \xi = (y - \hat{y})$$

To evaluate the full predictive densities, this score is then averaged over all target quantiles, from 0.01 to 0.99, for all time periods throughout the forecast horizon. The lower the pinball score, the better the forecast.

a) Why Quantile Forecasting?: In today's competitive and dynamic environment, more and more decision making processes in the power and energy industry are relying on probabilistic forecasts. The applications of probabilistic energy forecasts
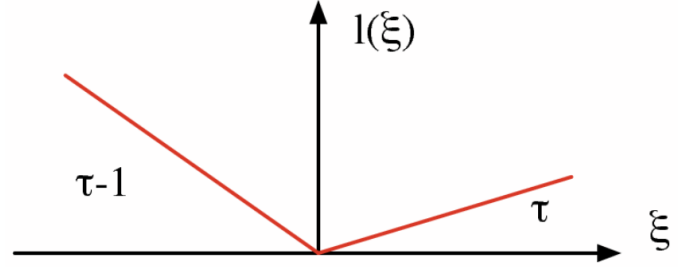

Fig. 1: The tilted loss function for 50th, $\tau = 0.5$, and 75th, $\tau = 0.75$, quantile [3].

spread across planning and operations of the entire energy value chain.

In our own evaluation of forecasts, we further use some well-established point error metrics, for the simple reason that we generate our quantile predictions starting from a point prediction. The measures are the Mean Average Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

The MAE and RMSE measures are the two most common scale-dependent errors, meaning that the residuals ξ_i (for ith observations) are on the same scale as the data. Hence, MAE and RMSE are in units of Fahrenheit or Mega Watt for our data set.

$$\text{MAE} = \mathbb{E}[\xi_i]$$

$$\text{RMSE} = \sqrt{\mathbb{E}[\xi_i]}$$

As a percentage error the MAPE measure is scale-independent. In our case it is useful for giving an immediate sense of the relative magnitude? of the error.

$$\text{MAPE} = \mathbb{E}\left[\left|\frac{100\xi_i}{y_i}\right|\right]$$

The measure is undefined for $y_i = 0$. Fortunately, in our dataset all values are several integers larger than zero for both load and temperature series so that the measure is neither undefined or affected by extreme values.

C. Data Cleaning

An annoying feature of the dataset is that the timestamps are not saved in the international ISO 8601 standard, but as "MMddYYYY H:m" without leading zeros for both days and months. Fortunately, the dataset was provided continuously without gaps and therefore the problem could be easily solved by hard-coding the first and last datetimes and using these to generate the needed sequence of datetimes.n

D. Temperature Data Selection

We use the average temperature of all weather stations at any hour i as the basis series for our temperature forecasts (Figure 3a). As the basis for the load forecasting we use the average over all the temperature stations provided in the dataset. In other words, we use the average district temperature to predict the district utility load.

[?] temperature highly correlated → use average however, did not use temperature differences: would have led to 25x25 more features

In another approach we compute the cross-correlations of the 25 weather station's series (Figure 2). These suggest that the district temperature can be explained on average to over 90% by the first series.
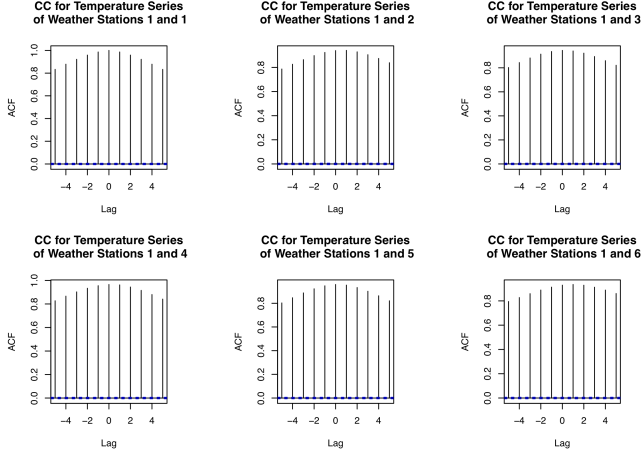

Fig. 2: Cross Correlation Plots of Temperature Station 1 Series with Series of Temperature Stations 1-6. Cross Correlations for Stations 7-25 ommitted for convenience of display.

A temperature of 60 degrees Fahrenheit would therefore allow for an error of maximum 6 degrees of Fahrenheit or 3 degrees Celsius respectively. As we will see later in this report this error is negligible when taking into account the inaccuracy of/the error introduced in the temperature prediction.
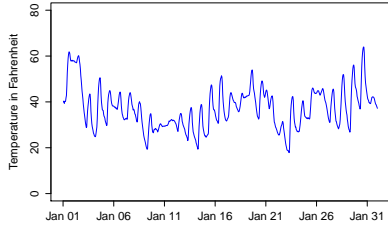
(a) Average Zonal Temperature during July 2011
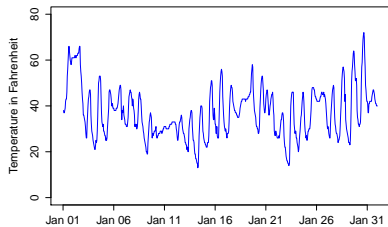
(b) Temperature of 1st Weather Station during July 2011

IV. FEATURE SELECTION

In the following paragraphs we explain the methodology we use to obtain our final feature selection for both temperature and load forecasting. The choice of features is limited to calendar features for temperature forecasting, because the only additional data allowed in the competition are Federal US holidays and the forecasting horizon is set to 1 month ahead.

A. Time Lags

First of all, we want to find out which previous values can be useful in a prediction. Given a time horizon of 1 month we would like to forecast the whole month using only one model for reasons of convenience. This circumstance induces a time lag of minimum 28 days for the month of february. For simplicity of implementation we define the same minimum lag of 31 for every month of the year. In order to choose the lag variables we interpret the cross-correlation of observations y_i and y_{i-k}, for different hourly lags $k = 1, 2, ..$, of both temperature and load time series in (Figure 4). This cross-correlation of observations is known in statistics as Autocorrelation and demonstrates the similarity of observations in a series as a function of the time lag between them. Using the empirical mean and standard deviation of our target series $Y = \{y_1, y_2, \ldots, y_i, \ldots, y_N\}$ we can compute the autocorrelation for lag k with:

$$\mathcal{R}(k) = \frac{\mathbb{E}\left[(y_i - \mu)(y_{i-k} - \mu)\right]}{\sigma^2}$$

As can be seen in both figure 4a and 4b the highest autocorrelations occur with a frequency/periodicity of 24 hours. Therfore we choose 24 hours as the basis constant for selecting the time lags of our predictions. The average temperature series in the former figure displays less variance in the correlations than the latter because the hourly temperature is less extreme due to zonal averaging.
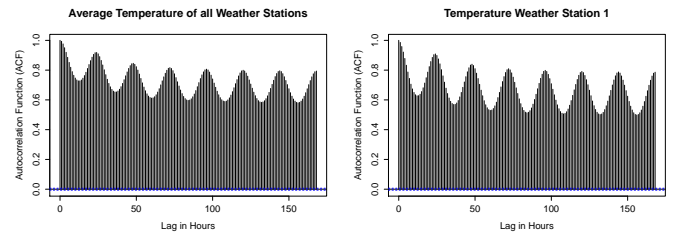
(a) 7 days Autocorrelation   (b) 7 days Autocorrelation

Fig. 4: ..

Figure 5 shows the autocorrelation function of the load series for different maximum lags. As can be seen the load correlations also have a periodicity of 24 hours (5b). The other subfigures show that after an initial exponential decrease the correlations decrease near linearly. Since the type of day has an influence on the utility load (Section IV-B), we set the basis for time lags to $7 \cdot 24$ hours for the load and the minimum time lag for monthly load forecasts therefore to 35 days.

(a) 7 days Autocorrelation
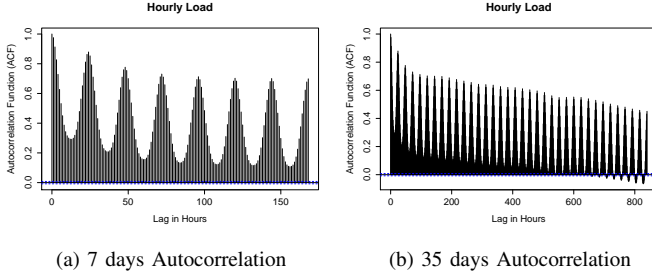
(b) 35 days Autocorrelation

Fig. 5: Autocorrelation Function Estimates for hourly load data in MW for different maximum lags

B. Calendar Features

Figures 4 and 5 have shown both temperature and load series to possess daily frequency. We therefore extract the hour of the day from the datetimes and use it as a feature.

Figure 6 demonstrates the seasonal change of temperature and load throughout the year. In order to capture the seasonality in our models for both temperature and load, we use the time of the year as a feature. We indicate the progress of the current year with a sequence TOY $\in [0, 1]$ [?]. We dropped the original categorical variable indicating the present month in favor of the time of year feature (validated by experiment.)

(a) Autocorrelation Estimates for daily average district temperature.

(b) Autocorrelation Estimates for daily average district load.
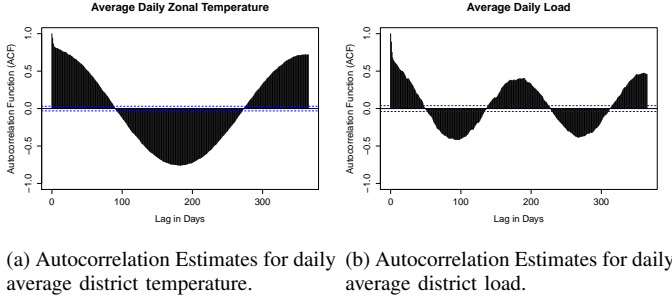
Fig. 6: Autocorrelation Function Estimates for daily averages of both district temperature and load for 1 year.

The features we discussed so far are proven useful for both temperature and load forecasting later in this report (Section ??). The nature of the load data allows us to proceed with the extraction of another feature, the day of the week as introduced by [4]. Energy consumption does not only depend on the climatic circumstances, but also directly on the calendar events, because consumption patterns vary between the different days of the week, weekends and holidays due to human activity (find source). The energy utility data does not show clear patterns when probing (Figure coming up). Nevertheless, as proposed in [2] we use different approaches of marking the days of the week, numbering the days of the week from 1 (Sunday) to 7 (Saturday) and assigning weekdays and weekend days with 1 and 2 respectively.

The only external data source allowed for the GEFCom load forecasting track are U.S. federal holidays for the time span of the provided data. This gives us the possibility to addi-

tionally detect holidays and assign them another integer (8) in the approach of consecutively numbering the days of the week.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Holiday
Weekdays vs -end	2	2	2	2	2	1	1	-
Weekdays	2	3	4	5	6	7	1	-
	2	3	4	5	6	7	1	8

TABLE I: My caption

V. FORECASTING METHODS

(find paper using generative additive models for forecasting) In this section we briefly review the forecasting models used during this project. On the one hand we use a generalization of the Linear Model (LM), the Generalized Additive Model (GAM), and on the other basic implementations of established machine learning methods such as Neural Networks (NN) and Random Forests(RF). Due to the limited scope of the project we did not use our own implementations, but rather those that have been made available through R packages.

instead of linear predictor refer to univariate response variable

A. Linear Model (LM)

We use the ordinary least squares (OLS) linear regression model as a benchmark method to interpret the prediction power of the other methods. OLS assumes the residuals to follow a normal distribution $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$. The linear predictor is therefore the sum of a linear combination of covariates and Gaussian noise and the predictor $y = \phi^T \mathbf{x} + \epsilon$. The vector of covariates $\mathbf{x}$ contains the pre- selected features, i.e. the temperature variable and the calendar effects, as well as the bias term x_0 and ϕ represents the vector of coefficients that are to be learnt with least squares approximation.

B. Generalized Additive Model (GAM)

Why do we use GAM? The standard linear regression model we use is a generalized linear model with a gaussian distribution and an identity link function. For the general additive model [5] we employ we will not change the link function, since we are going to use a Gaussian distribution here as well. However, generalized additive models allow a generalization on top of that of the incorporation of other distributions as in Generalized Linear Models, that of *nonlinear predictors*. Hence, assuming a Gaussian error model the linear predictor turns into a nonlinear predictor with a combination of smooth functions of the predictor variables: $y = s_1(x_1) + s_2(x_2) + \cdots + s_p(x_p) + \epsilon$, where the $s(x_i)$ are smooth functions of the chosen features. Note that not all covariates need to be wrapped in smooth functions.

Since the nonlinear functions are smooth they can be estimated by penalized regression in a spline basis [6], [7]:

$$s_i(x) = \sum_{j=1}^{k_i} \beta_{i,j} \varphi_j^i(x)$$

Here k_i is the dimension of the spline basis and φ_j^q are the corresponding spline functions. The default splines for GAMs in the **mgcv** package are thin-plate splines [8] where $\varphi(x)$ is of the family of radial basis functions. The focus of this project was on forecasting workflow and feature selection, so that different spline specifications have not been tried out.

Show model output and how degrees of freedom are chosen?

C. Neural Network (NN)

Neural networks are powerful machine learning algorithms. They exist in many varieties, with feedforward neural networks being the most basic. In our predictions we use a simple feedforward neural network with one hidden layer with a varying number of hidden units (from now on referred to as **hunits**) and a linear output unit. Such a network can be easily used for regression using the R package *nnet* [9].

D. Random Forest (RF)

Random forests [10] are a powerful ensemble method for capturing nonlinear dependencies that does not require significant tuning. They employ decision trees which are very popular in machine learning. To avoid the effects of overfitting a bunch of random decision trees are automatically created and the mode of the probability density of the predictions is returned as the point predictions. We use the *randomForest* R package [11] and try the algorithm for a varying number of trees. We will refer to this variable in the following as **ntrees**.

A decision tree can process continuous and discrete covariates, as is the case with our features, while providing a continuous response. Chapter 15 of [12] provides the reader with a detailed introduction to the subject. *To do this, each simple tree gives its prediction, and these are then aggregated using the mean of the individual predictions. Each tree model assigns a class to the new vector of covariates, based on the recursive partitioning. Then, the prediction is the mean value of the responses that correspond to this class* [6].

VI. PIPELINE

A. Creation of Features

B. Crossvalidation

C. Monthly vs. Weekly

D. Creation of Quantiles

We use the residuals $\xi = y - \hat{y}_i$ created during training to create the desired quantile forecast. Error shoud ideally have zero mean using model point predictions as the mean.

VII. ANALYSIS

What combination of features and models for temperature and load provide us with a good prediction accuracy with respect to Gefcom leaderboard?

predtrain vs no pred train (GAM: LoadFormula 1, tempFormula 1) $\rightarrow$: $weekly pred train has better score, but only one position on average$

GAM: combined weekly temperature predictions worse than monthly prediction, model with DLAG slightly worse than without

the first week is the worst on average Combined weekly predictions better for GAM model! Does DLAG have positive effect?

A. Temperature Modeling

1) Data Processing: average temperature vs. principal component

2) Effect on Load Prediction: Effect of temperature on load prediction evaluated using different methods:
Mean over past years (yearly lag), LM, GAM, NN, RF vs. true temperature

3) Evaluate Results of weekly vs. monthly Temperature Prediction: plot MAPE & PINBALL scores for different methods over load training + CV period in a 1x2 plot of the form:
monthly scores   weekly scores

B. Load Modeling

1) Evaluate the Influence of the Lag on Load Prediction: set the basis by plotting MAPE & PINBALL scores by week over w1, w2, w3, w4; one curve for every month in CV
do this for every method
as well as a comparison of the best performing configuration of every method among each other

2) Evaluate Performance for different Method Configurations: Use temp method that provides best score as shown in Temperature Modeling Section

Different GAM formulas:
plot MAPE & PINBALL scores for all GAM formulas over CV period in a 2x2 plot of the form:
monthly load with monthly temp   monthly load with weekly temp
weekly load with monhtly temp   weekly load with weekly temp

Different NN hidden units:
plot MAPE & PINBALL scores in 2x2 plot

Different RF ntrees:
plot MAPE & PINBALL scores in 2x2 plot

3) Compare Performance of different Methods: choose best scoring configuration for every method and plot the results in one 2x2 plot

VIII. CONCLUSION

The conclusion goes here.

REFERENCES

[1] T. Hong, J. Wilson, J. Xie, and A. Member, "Long Term Probabilistic Load Forecasting and Normalization With Hourly Information," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 456–462, 2014.

[2] R. J. Hyndman and S. Fan, "Density Forecasting for Long-Term Peak Electricity Demand," *IEEE Transactions on Power SystemsHyndman, R. J., & Fan, S. (2010). Density Forecasting for Long-Term Peak Electricity Demand. IEEE Transactions on Power Systems, 25(2), 11421153. doi:10.1109/TPWRS.2009.2036017*, vol. 25, no. 2, pp. 1142–1153, May 2010. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5345698

[3] I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola, "Nonparametric Quantile Estimation," *Journal of Machine Learning Research Nonparamteric Quantile Estimation*, 2005.

[4] T. Hong, M. Gui, M. E. Baran, S. Member, and H. L. Willis, "Modeling and Forecasting Hourly Electric Load by Multiple Linear Regression with Interactions," p. 4, 2010.

[5] S. N. Wood, *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2006.

[6] R. Nedellec, J. Cugliari, and Y. Goude, "GEFCom2012: Electric load forecasting and backcasting with semi-parametric models," *International Journal of Forecasting*, vol. 30, no. 2, pp. 375–381, Apr. 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169207013000800

[7] M. Clark, "ADDITIVE MODELS." [Online]. Available: http://www3.nd.edu/~mclark19/learn/GAMS.pdf

[8] S. N. Wood, "Thin-plate regression splines," *Journal of the Royal Statistical Society (B)*, vol. 65, no. 1, pp. 95–114, 2003.

[9] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[10] L. Breiman, *Random Forests*, 2001.

[11] A. Liaw and M. Wiener, "Classification and Regression by randomForest." *R News: The Newsletter of the R Project*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://cran.r-project.org/doc/Rnews/

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.