Can we predict our mood based on cellphone data?

Group 57: Xinyu Hu (2691175), Christophe Meijer (2585818), and Andrei Udriste (2712179)

Vrije Universiteit Amsterdam

1 Introduction

A healthy environment is often determined by the mental health of individuals. Mental health stimulates general motivation in achieving life goals and can even benefit the economy [2]. Individuals are more conscious about their mental health than before as the internet is increasingly used to seek information about medical issues. Not everybody is aware of problems they got and mental health problems caused by unconscious stress can occur [3]. Much of the stress and straining of people remains hidden [12]. A smartphone is often used as a coping mechanism to handle different kinds of stress. Findings in a Chinese study revealed that both coping style and emotional intelligence mediated a link between smartphone addiction and psychological abuse [11]. Another study substantiated the same conclusion: Neuroticism and negative coping style mediated the relationship between childhood psychological maltreatment and smartphone addiction [6]. However, this seems like a bad coping mechanism. Smartphones can be a cause of sleep deprivation, depression and anxiety [5]. The increased need for touch, fear of missing out leads to more problematic mobile usage. This is dangerous as it can cause a vicious circle where people are being enslaved to their phone with increased negative emotions. A couple of solutions have already been provided. A controversial one is to transform Judas into Jesus. Smartphone applications are being developed to prevent anxiety and increasing mental health. However, few have been evaluated in terms of usability and or even clinical effectiveness. A study who did, provided hopeful results in which they showed that the created app was highly and positively rated by youth and providers [10]. This study did not distinguish the influence of other apps on the mood of the participants. It could be possible that the influence of the use of social media apps interferes with this effect. Another study showed that increased usage of another developed app decreased anxiety and depression [1]. Once again, other influences were not taken into consideration. Applications like this could potentially assists depressive clients in difficult situations. Based on measurements, a model then determines if a supportive intervention is triggered. Previous data is used to make an accurate prediction. A precise prediction can contribute to increased quality of therapy or prevention of mental diseases. Smartphones are popular devices, capable of processing a lot of information. They include many features such as games, access to the Internet and social networks, messaging,

weather applications, multimedia in addition to their use of communication. In order to pinpoint what causes mood changes a clear statistical analysis needs to be made. In this paper we make an attempt to predict the mood level of participants by using mobile phone usage data and simultaneously compare different statistical models.

2 Data and method

2.1 The Data

We utilized a big dataset provided by Vrije Universiteit van Amsterdam. The domain from which the dataset originated is the domain of mental health. The dataset contains ID's, reflecting the user's measurement originated from. Furthermore, it contains time-stamped pairs of variables and values. The dataset consists of 26 patients who reported their mood level for 3-4 months. The clients had the possibility to report their mood level on a scale from 1 to 10. They were also asked to report valence and arousal on a scale from -2 to 2. Figure 1 gives an insight in how many datapoints are collected per patient and per attributes. As you can see the amount of datapoints are shown per patient. Patient 14.01 collected the most datapoints and 14.15 the least. In the second histogram duration of screentime was captured the most and appCat weather the least.

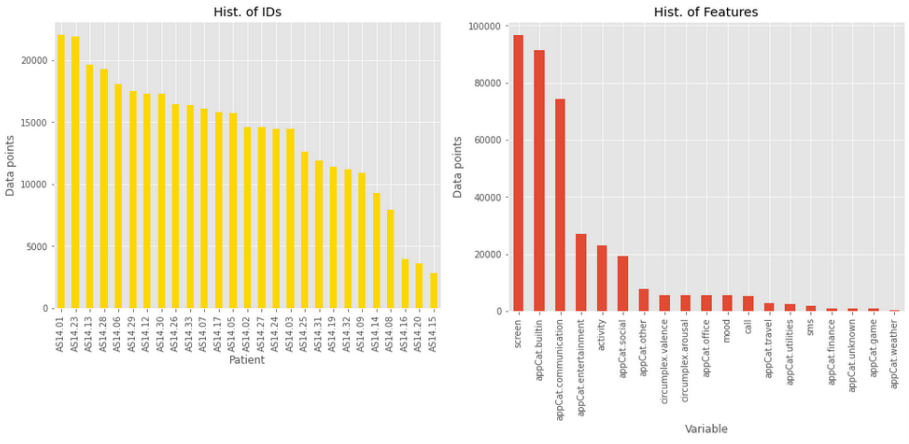

Fig. 1. Histograms of IDs and features

2.2 The approach

For analyzing the data, we tried to preprocess the data. A lot of variables weren't filled in completely, leading to gaps in our dataset. Based on some descriptive statistics we cleaned the dataset. Furthermore, we decided to analyze the data on a personal level, because this is more insightful than considering every datapoint

as independent from the other datapoints. One can imagine that the mood at 10am is a good predictor of the mood at 12 am. For every model we created a training set and a test set. The training set was approximately 80 percent of the whole dataset and the test was based on the other 20 percent. However, the size of the datasets and the train/test split ratios can greatly affect the outcome of the models [8]. A good and allround acceptable ratio is the 80-20 split we used. With less training data, the parameter will have greater variance, with less testing data, the performance statistic will have greater variance. The train/test split ratio can vary for different models.

2.3 Models

Baseline models In this paper we differentiated three baseline models. The first baseline predicted the mood of the next day by saying it's the same as the previous day. The second baseline calculated the average of a time-framed window. This window consists of mood values from 5 sequential days and the average is calculated. This will be the prediction of the next day. The sliding window moves one up and the new average will be calculated, and so on. The third baseline is called the random probability baseline. The values of the previous day are taken into consideration and a value is picked from a uniform distribution based on a chosen probability P. The values from the uniform distribution range from -0.5 to 0.5. If there is determined that a value is picked from this distribution it will be added to the value of the previous day and determine the value for the next day. This process will be repeated. The reason the uniform distribution ranges the way it does, is because any higher values will deviate the prediction values too much and lower values will deviate the prediction too less.

Regression Models Furthermore we utilized variant regression models to predict the mood levels for each patient. The relationship between the mood level (outcome variable) and screen (independent variable) are represented by coefficients. Results will be compare with other similar regressions like Ridge Regression, Lasso Regression, Elastic Net Regression which is measured by mean squared error. However, each regression model generated the same plots. It shows that the models have the same predictions in case of same input datasets. The below is the plot for a patient.

ARIMA We utilize the ARIMA model. Autoregressive integrated moving average (ARIMA) is one of the popular linear models in time series forecasting during the past three decades [13]. ARIMA can be best understood when the three most important components are differentiated. Autoregression (AR) covers the part of the model which explains a changing variable that regresses on its own lagged, or prior, values. The Integrated (I) has to take care of the time series to become stationary. Data values are being replaced by the differences between the data values and the previous values. A model is stationary when it shows a certain constancy over time (no trend). Moving average (MA) incorporates the

dependency between an observation and a residual error from a moving average model applied to lagged observations.

LSTM We utilize the LTST (Long short-term memory) model. This is an artificial recurrent neural network. A LSTM unit consists of four components. A cell, an input and output gate and a forget gate. The three gates supervise the flow of information into and out of the cell and remember values over arbitrary time intervals. LSTMs are perfect for time series, since there can be lags of unknown duration between important events in a time series.

Xgboost Xgboost provides a way to measure the importance of each attribute. Typically, the importance provides scores that indicated the usefulness or value of each feature in constructing an enhanced decision tree in the model. The more attributes used for the decision tree, the higher their relative importance. This importance is calculated explicitly for each attribute in the dataset so that the attributes can be ranked and compared to each other.

2.4 Performance measurement

For calculating the accuracy between the baseline models we used MAPE for measurement. The mean absolute percentage error (MAPE) is a statistical measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values. We utilized the Root Mean Square Error (RMSE) as a comparison and performance measure. The root mean square error (RMSE) has been used as a standard statistical metric to measure model performance in meteorology, air quality, and climate research studies [4]. The RMSE can be defined as the average distance between predicted and observed values.

3 Results

3.1 Preprocessing

Data cleaning Mood is a temporary state of mind or feeling. This feeling is mostly stable during the day in healthy people [7]. Based on this paper we performed our first step: To resample the timestamps from hours to days. The second step was to adjust the attributes call and sms. When there was a call or sms made the number 1 was noted and the number 0 if otherwise. The dataset stored the zero's as NAN values. These had to be switched back to zeros again. The third step was to eliminate every NAN value from mood. The reason for these NAN values were because patients didn't fill in the surveys completely. Most of the other surveys were not filled in completely. This was leading to too many gaps to make reliable analysis. Therefore we decided to create a function

that deletes all the patient's attributes that have more than 30 percent of their data missing. There was only 1 patient who neglected to fill in the surveys. By deleting this patient ('AS14.26'), attributes like activity, screen and appCat.builtin passed the limit of 30 percent and could be included. The rest of the NAN values in attributes was filled in with either the median of the attribute or the mean of the attribute. The mean was used when there were not many outliers (screen, call). The median was used when there were plenty of outliers [9] An example of how an attribute looks like before preprocessing is shown in figure 2. An example of how an attribute looks like after preprocessing is shown in figure 3. The fifth step was to cut extreme outliers. In the last step we deleted highly correlated attributes. In figure 4 a correlation matrix of the attributes is shown. A high correlation can have a bad outcome for a predicting model so valence got deleted from the dataset.

Fig. 2. Example of the attribute activity before preprocessing. The amount of activity on the x-axis and the time on the y-axis

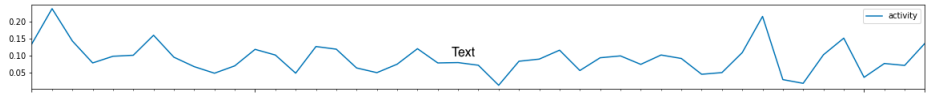

Fig. 3. Example of the attribute activity after preprocessing. The amount of activity on the x-axis and the time on the y-axis

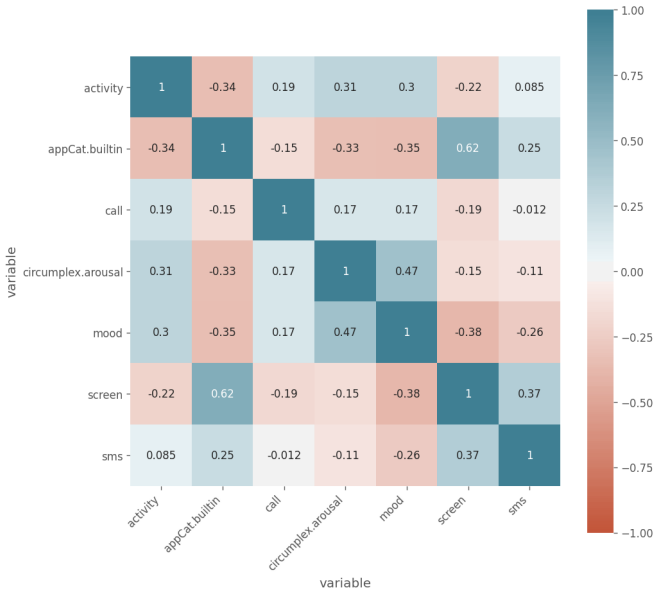

Fig. 4. Correlation matrix of different attributes. Mood correlates highly with valence.

3.2 Models

Baseline models In figure 5 the description of all baselines are displayed. Here we see the MAPE values for each of the baselines. MAPE value for moving average was the highest with 8.0, base was second with 6.92 and prob ended last which had a MAPE value of 7.0.

	MAPE_base	MAPE_ma	MAPE_prob	MAD_base	MAD_ma	MAD_prob	corr_base	corr_ma	corr_prob
count	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	8.0	26.000000	26.000000
mean	7.072308	6.978846	9.280769	-0.074615	0.092308	-0.090000	0.0	0.060000	-0.138462
std	3.359536	3.794473	4.165494	0.398228	0.293807	0.596081	0.0	0.314019	0.348042
min	2.750000	2.140000	4.100000	-1.040000	-0.490000	-1.740000	-0.0	-0.610000	-0.640000
25%	4.175000	4.875000	6.507500	-0.255000	-0.127500	-0.437500	-0.0	-0.050000	-0.450000
50%	6.805000	6.345000	8.145000	-0.015000	0.115000	0.005000	0.0	0.045000	-0.215000
75%	8.790000	8.155000	11.032500	0.175000	0.235000	0.287500	0.0	0.232500	0.115000
max	15.180000	20.140000	22.060000	0.460000	0.620000	0.840000	-0.0	0.750000	0.450000

Fig. 5. Accuracy of all the baselines

Regression models As it can be observed in Figure 5 all the Regression models look the same. None of the regression model outperforms.

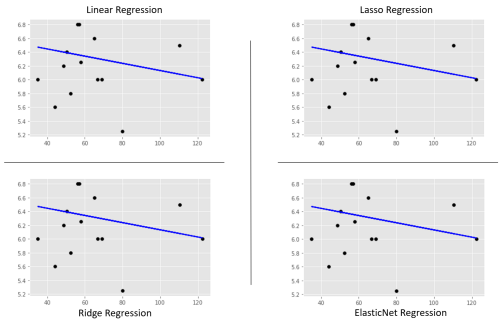

Fig. 6. Top left: Linear regression. Top right: Lasso Regression. Bottom left: Ridge regression. Bottom right: Elastic Net Regression

In figure 6 the importance of features is calculated with the xgboost method. The F-score is displayed on the x-axis. This represents the importance of an attribute. The average F-score of each model is displayed for 50 repetitions of the fivefold cross-validation (CV) carried out in the training set. Screen predicts the value of mood the most where sms predicts the value of mood the least.

We used every attribute for the calculation of mean squared error accuracy (40.11 percent). The less attributes we use for this calculation the less it predicted the mood.

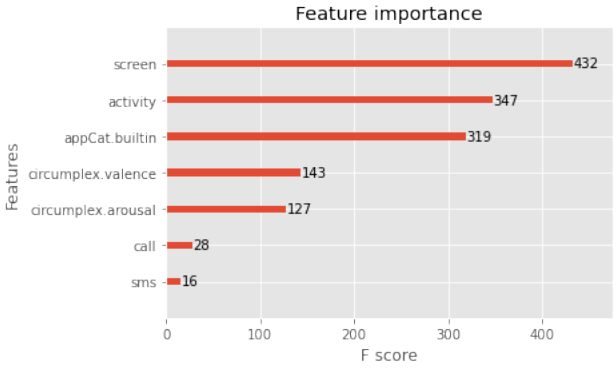

Fig. 7. The importance of features

ARIMA The first temporal mode used to predict the value of the mood was the ARIMA model. This represent one of the most popular approaches to predict temporal data, especially when we are talking about univariate temporal analysis. Because for the given data we had a number of patients that each had a different mood we had to use a grid search approach to find an ARIMA model that will best fit all of the patients. After implementing the grid search we obtained that the best value for p=2, for d=0 and for q=0. But there is another approach to finding the hyper-parameters of a ARIMA model, mainly manual search, this means trying every value and see what is the best fit over the predicted data, if we try to do that we obtain the following values: p=2, d=1 and q=0. After implementing the hyper-parameters obtained by grid search and manual search we obtain the plot presented Fig 7.

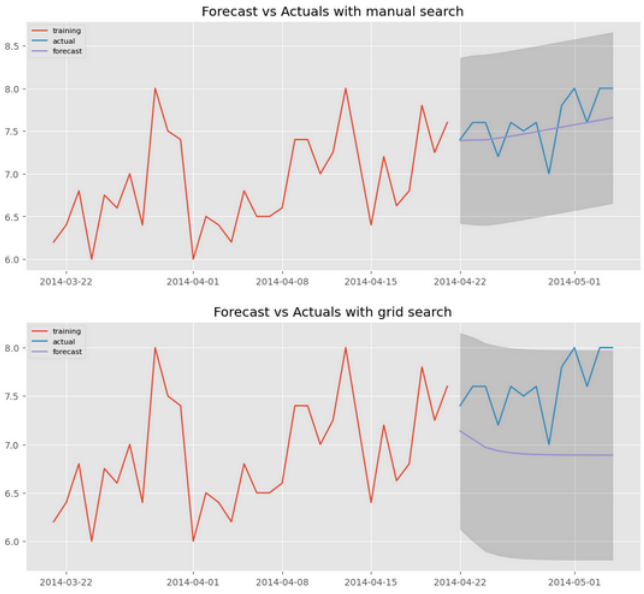

Fig. 8. Tow plots showing the difference between the grid search approach (bottom graph) and the manual sear (upper graph) for the selection of the hyper-parameters for the ARIMA model

As it can be observed the manual hyper-parameter have a much better fit than the grid search one, which would indicate that our grid search has produced wrong results. But this is not true, because we did a grid search over all the patient, while in this case we only try to fit the hyper-parameters for one specific patient. Because of this the manual search hyper-parameter are not an option, especially since we have a high number of patient, so we will have to rely on the hyper-parameter obtained from the grid search.

	MAPE	MAD	corr
count	26.000000	26.000000	26.000000
mean	7.134615	-0.021923	0.111923
std	3.558477	0.378048	0.321770
min	2.350000	-1.140000	-0.490000
25%	4.500000	-0.197500	-0.120000
50%	6.700000	-0.040000	0.060000
75%	9.027500	0.280000	0.337500
max	16.540000	0.560000	0.700000

Fig. 9. Accuracy estimations for the ARIMA model

If we have a look at Fig 8 we can see the accuracy result obtained from after calculating the MAPE for all the patients and averaging all those results we obtain a MAPE of 7.13. This is a very interesting result, mainly because the baseline is better than the ARIMA model. Which is a surprise, since we would expect that the ARIMA model would be better than our implemented baseline, but we can see that this in not the case. Another observation would be that this does not apply for all the baselines, this only applies for the normal baseline and the moving average baseline, both having a MAPE smaller than 7.13, but the probability baseline has a MAPE of 9.28, which make's it worse.

LSTM After implementing the ARIMA model and seeing that we haven't got a better approximation of our data we decide to also implement an LSTM Neural Network. Hoping that this will offer us a better prediction for the data that we have. This has been done using mainly the library Pytorch, a library dedicate for neural networks (NN). For this task we used the mood as our input variable in the hopes that we will obtain a prediction for it. The shape of the NN is 1/100/1, this means that we have one value as an input, 100 neurons in the hidden layer and one value as the output. The NN has been trained using a sliding window method, where you group the days into a tuple, where the first element is the days that you have and the second element is the day that you try to predict. Because we try to train the model to predict the mood for all patents we decide to train on the training data for all the patients, and we train it for a total of 50 epochs. After the 50 epoch we have a loss of 0.0994804502.

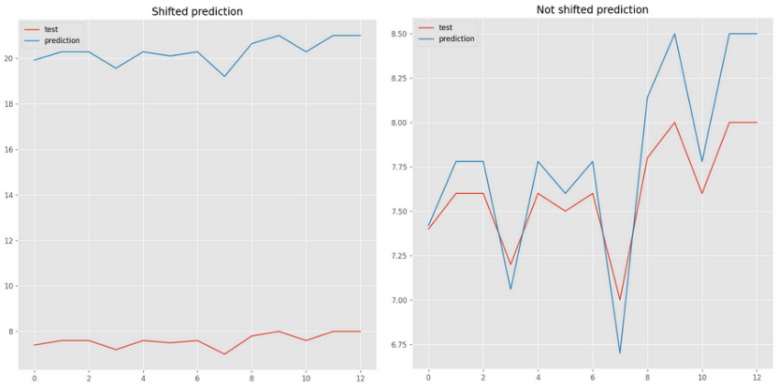

Fig. 10. The predictions for one patient obtained using the LSTM NN. Left graph the shifted data. Right graph corrected data

In Fig 9 we can observe that LSTM produces a pretty good prediction of the next day, but there is also a very big problem we can see a shift in the data. If we look at the plot from the left we can observe just how big the shift in the data is, but there is also a good part, the predicted values are just shifted, but it looks like the prediction are pretty accurate. To remove the shift in the date we can just decrease all the values with 12.5, and in this case we obtain the plot from the right side of Fig 9. Here we can observe that the predictions of the LSTM look much more closer to the true values. Now is time to do some statistical tests to see how close the predicted numbers are to the actual numbers.

	MAPE	MAD
count	26.000000	26.000000
mean	12.993077	-0.295769
std	6.163004	0.387641
min	5.780000	-1.300000
25%	8.562500	-0.427500
50%	12.455000	-0.255000
75%	16.127500	-0.132500
max	33.930000	0.410000

Fig. 11. Accuracy estimations for the LSTM NN

If we look at Fig 10 we can observe the accuracy estimators obtained for the LSTM model, and we can see that they are even worst than the ARIMA model. This fact is very surprising since we can clearly observe from the plots that the predicted values are very accurate. But apparently the difference between them is misleading and the LSTM model has introduced some lag in the model, signified also by the shift in data, that affects our prediction accuracy. Another observation would be that the LSTM model performs the worst out of all the models, by an accuracy standpoint, the MAPE for it being almost double

than the one for the best performing model (the base baseline). The last observation would be that even tough the lstm perform the worst from an accuracy standpoint, is the best one from a visual standpoint, the plot obtained using the LSTM values is the closest one to the graph obtained from the real data.

4 Conclusion

We performed a multitude of different statistical methods on a dataset that includes the data of observations of own mood from patients and their mobile phone usage. In conclusion, we find that the LSTM model fitted the test data the best on the graphs, but when it came down to statistical accuracy it scored the worst. Baseline MA scored the highest on accuracy measurement. Then Baseline base, Arima, Baseline probability and LSTM. All the models are okay dependent on what you're looking for. To make a better forecasting model we assume that obtaining more meaningful attributes than mood might contribute to the forecast more intensively can boost prediction performance as well. For example, the influence of the days can be taken into consideration. One can imagine that a Monday has a bad influence on the mood, may be more than some of the applications do. Another option is to improve ARIMA models by creating a model for each individual. However, this can be problematic for big datasets. The LSTM model can be improved by creating more neurons and more epochs. However this is more expensive and may overfit. In the future, we aim to build models that are able to predict the mood level of subjects more accurately and find ways to select important features more reliably.

References

1. Bakker, D., Rickard, N.: Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: Moodprism. Journal of Affective Disorders **227**, 432–442 (2018)
2. Becker, D., Bremer, V., Funk, B., Asselbergs, J., Riper, H., Ruwaard, J.: How to predict mood? delving into features of smartphone-based data (2016)
3. Brosschot, J.F., Verkuil, B., Thayer, J.F.: Conscious and unconscious perseverative cognition: Is a large part of prolonged physiological activity due to unconscious stress? Journal of psychosomatic research **69**(4), 407–416 (2010)
4. Chai, T., Draxler, R.R.: Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscientific model development **7**(3), 1247–1250 (2014)
5. Demirci, K., Akgönül, M., Akpinar, A.: Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. Journal of behavioral addictions **4**(2), 85–92 (2015)
6. Liu, F., Zhang, Z., Chen, L.: Mediating effect of neuroticism and negative coping style in relation to childhood psychological maltreatment and smartphone addiction among college students in china. Child Abuse & Neglect **106**, 104531 (2020)
7. Polak, M.A., Richardson, A.C., Flett, J.A., Brookie, K.L., Conner, T.S.: Measuring mood: considerations and innovations for nutrition science. Nutrition for brain health and cognitive performance pp. 95–122 (2015)

8. Rácz, A., Bajusz, D., Héberger, K.: Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. Molecules **26**(4), 1111 (2021)
9. Somasundaram, R., Nedunchezhian, R.: Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. International Journal of Computer Applications **21**(10), 14–19 (2011)
10. Stoll, R.D., Pina, A.A., Gary, K., Amresh, A.: Usability of a smartphone application to support the prevention and early intervention of anxiety in youth. Cognitive and behavioral practice **24**(4), 393–404 (2017)
11. Sun, J., Liu, Q., Yu, S.: Child neglect, psychological abuse and smartphone addiction among chinese adolescents: The roles of emotional intelligence and coping style. Computers in Human Behavior **90**, 74–83 (2019)
12. Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., Campbell, A.T.: Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing. pp. 3–14 (2014)
13. Zhang, G.P.: Time series forecasting using a hybrid arima and neural network model. Neurocomputing **50**, 159–175 (2003)

References