A COMPREHENSIVE ANALYSIS OF UNIVERSITY FITNESS CENTER DATA UNCOVERS INTERESTING PATTERNS, ENABLES PREDICTION,

Retz Mahima Devarapalli, Uppalapati.Sireesha

ASSISTANT PROFESSOR, DEPT OF IT, VIGNAN'S LARA INSTITUTE OF TECHNOLOGY AND SCIENCE, VADLAMUDI, ANDHRA PRADESH 522213.

MCA STUDENT, VIGNAN'S LARA INSTITUTE OF TECHNOLOGY AND SCIENCE, VADLAMUDI, ANDHRA PRADESH 522213.

Abstract:

Data is progressively being utilized to make regular day to day existence simpler and better. Applications, for example, holding up time estimation, traffic prediction, and stopping search are genuine instances of how data from various sources can be utilized to encourage our day by day life. In this investigation, we consider an under-used data source: college ID cards. Such cards are utilized on numerous grounds to buy food, permit access to various zones, and even gauge participation in classes. In this article, we use data from our college to examine utilization of the college wellness focus and manufacture an indicator for future visit volume. The work makes a few commitments: it exhibits the lavishness of the data source, shows how the data can be utilized to improve understudy administrations, finds intriguing patterns and conduct, and fills in as a contextual investigation delineating the whole data science process.

Index Terms—Pattern analysis, Modeling and prediction, Data mining, Machine learning, Time series analysis, Computer applications miscellaneous.

I. INTRODUCTION

WASHINGTON State University is a land-award college with an understudy populace of approximately 20,000 on its primary grounds in Pullman. The Student Recreation Center (SRC) is among the most as often as possible visited grounds offices in the college. Passage to the SRC is observed using an official college ID card, the CougarCard; clients swipe the card upon section. Data collected from these card swipe exercises are rich and can be utilized to increase important bits of knowledge into grounds life and exercise practices. Be that as it may, little consideration has been paid in the past to this potential. One of the more extensive objectives of this is work is to exhibit how to tackle this likely by means of cautious analysis and to spike extra examinations nearby engaged information revelation.

The SRC is a well known spot to visit nearby, and alongside the fame, two particular kinds of necessities emerge normally. From the SRC administrators' perspective, realizing the use pattern of the offices is imperative to give agreeable understudy administrations. From the understudies' perspective, knowing future visit volumes is essential to have the option to maintain a strategic distance from the SRC when it is generally packed. This work targets tending to these twin needs all the while.

The methodologies we take line up with the two needs. To begin with, we use data mining procedures to reveal fascinating use patterns at the SRC dependent on verifiable data gathered from card swipes; this would give valuable data to the SRC directors to help with day by day tasks, for example, move booking or occasion arranging. Second, we consider the plausibility of utilizing these data to anticipate how packed the SRC will be for a given time span.

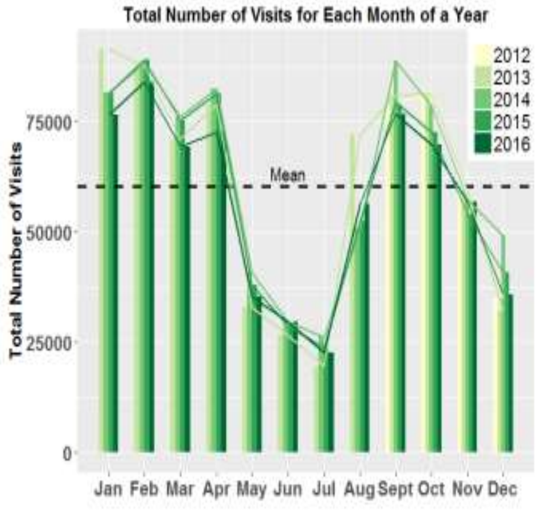


Fig. 1. This figure shows the total number of visits for each month of a year. The dashed line indicates the mean of all years.

In a fundamental report that showed up in a non-chronicled workshop paper [1], we had found fascinating utilization patterns by means of Exploratory Data Analysis (EDA) of card swipe data and utilized the found patterns to build up a choice tree model to make a six-way characterization: regardless of whether the SRC will be "swarmed," "swarmed," "reasonably swarmed," "somewhat swarmed," "not swarmed," or "practically unfilled" for a given time stretch. This article expands the fundamental work in a few different ways:

1) we gathered and investigated understudy card exercises for a more drawn out period;

2) we gathered an optional dataset—specifically, client profile data of the individuals who visited the SRC—to additionally understand understudy practice practices;

3) we created three distinctive prediction models to foresee future visit volumes at the SRC and considered which of the models plays out the best for our dataset (the investigation additionally empowered us to give understanding on the most proficient method to pick a reasonable model for other comparable data sets); and

4) we assembled an online application to show prediction results, which has helped SRC administrators in improving the nature of understudy administrations.

Our work is comparable in soul to a couple of existing applications that are based on information found by gathering and breaking down movement (frequently optionally related) data. The Orlando Undercover Tourist App [2] is one model. This versatile application assembles continuous office use data at the Orlando Disney World and evaluations holding up time with the goal that sightseers can design their visits all the more proficiently. Another model is the traffic-prediction versatile application Waze

[3] that gathers driver exercises and suggests the quickest courses. Maybe the nearest in similarity to the possibility of this article is the Popular Times work in Google Maps [4]. This area based help shows the most well known times for a spot looked on Google Maps where clients can decide the best time to go visit. One more model, the grounds centered stopping search portable application Kpark [5] screens how packed parking garages are on a college grounds by publicly supporting clients' reports on stopping accessibility and refreshing prediction maps intermittently.

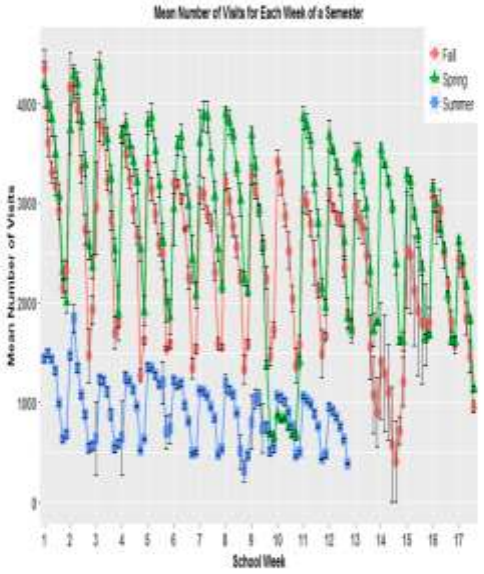


Fig. 2. This figure shows the mean number of visits for each week of a semester for the four academic years, 2012-13, 2013-14, 2014-15, and 2015-16. There are 17 weeks during Fall and Spring semesters and 12 weeks during Summer semesters. The standard error of the mean calculated across the four academic years is shown via the black bars.

Review of our discoveries. One objective of this article is to outline the data science process [6]. As far as data assortment, two arrangements of data—timestamp data from card swipes at the SRC and client profile data—were gathered from our University's CougarCard Center and Office of institutional examination (IR), separately. The gathered data was cleaned and further prepared. At that point, exploratory data analysis was performed to find fascinating patterns. General understudy practice patterns were found from the timestamp dataset. These incorporate yearly/month to month/every day frequencies of understudy visits to the SRC and pinnacle hours during a day. Also,we investigated understudies' profiles, searching for answers to addresses, for example, "Which sex or class level positions top in the recurrence of activity?" and "How does practice plan change with sex or class level?"

Making the following stride, utilizing information and knowledge picked up from the EDA, we develop a time series prediction issue and tackle it by building models for anticipating future visit volumes. Specifically, we fabricated and broke down three distinctive prescient models: an occasional guileless model[7], an autoregressive incorporated moving average(ARIMA)model [8], and a random backwoods (RF) model [9]. Both the occasional guileless models and the ARIMA model are generally utilized benchmark models to foresee time-series in the measurements network [10]. The RF model is a non-parametric model from the field of machine learning. Rather than the factual models, the RF model doesn't make any suspicions on the likelihood dispersion of the data [11], permitting us to legitimately gain from the data itself. Of the three, we found that the RF model yielded the best execution for our dataset as far as Root Mean Squared Error (RMSE), Mean Absolute Scaled Error (MASE), and Relative RMSE(RelRMSE).

Finishing the data science process, we fabricated an online application—a data item—and run the application live throughout the Spring 2017 semester (January to April 2017). The web application shows evaluated guest data. It was assessed by SRC representatives through May first, 2017.EDA helped us find a few fascinating patterns and practices with respect to clients of the SRC at our college. The greater part of these discoveries were straightforwardly utilized in the plan of the prescient models we assembled. A few, nonetheless, are of enthusiasm for their own privilege and can fill in as great exploration inquiries for future examination in related trains, for example, conduct or wellbeing sciences. We feature a portion of the key discoveries of our exploratory analysis beneath and prove them inSection 3.

- Student practice patterns have been reliable in the course of the last four scholastic years;

- Exercise schedules can be essentially influenced by occasions, get-aways, and school breaks;

- The SRC is visited more much of the time during weekdays than it is during ends of the week;

- Students will in general visit the SRC more regularly during Spring than Fall;

- Peak hours are in the early evening time during weekdays butin the early daytime during ends of the week;

- The length of stay is contrarily connected with appearance time—the later a client shows up at the SRC, the shorter they remain;

- Males visit the SRC more than females; and

- Freshman is the most successive client of the SRC contrasted with sophomores, youngsters, seniors, and graduate understudies.

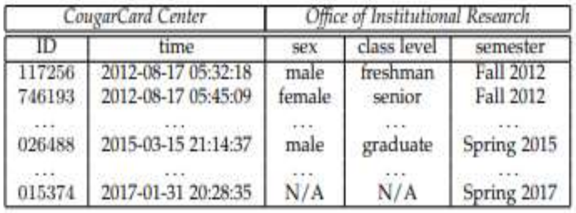
II.RELATED WORK

It comprises of future predictions of Students' temperament and scholastic exhibitions so the instructors can locate the correct procedure of handling and encouraging those Students with the goal that their scholarly presentation will get improved. The Student Recreation Center (SRC) is among the most habitually visited grounds offices in the college. Passage to the SRC is observed using an official college ID card, the Cougar Card; clients swipe the card upon section. Data reaped from these card swipe exercises are rich and can be utilized to increase significant bits of knowledge into grounds life and exercise practices. Be that as it may, little consideration has been paid in the past to this potential. One of the more extensive objectives of this is work is to show how to tackle this likely by means of cautious analysis and to spike extra investigations nearby engaged information revelation.

DATA

In this segment, we portray how our data was gathered and pre-handled. Sources. We originally accessed our essential data source, WSU's CougarCard Center, to gather CougarCard action data from across grounds. We separated more than 3 million wellness related records. The structure of this data comprises of timestamps for each card swipe when entering the SRC, organized as "year-month-date hour:minute: second." A randomized ID number was alloted to understudies to recognize people while protecting security. This gave us a preview of the wellness data at the SRC from August twentieth, 2012 to January 31st, 2017. One downside of this dataset was that it possibly permits us to know when understudies enter the SRC yet not when they leave. To conquer this, we introduced a transitory swipe-out framework at the leave entryway of the SRC from April first to June twelfth, 2016. This framework records the crash timestamp data in a similar arrangement as the swipe-in data. In any case, just a level of clients wiped out when they left, bringing about inadequate data. To cure this, we additionally recruited SRC representatives to physically tally the quantity of ways out for about fourteen days from May 31st to June twelfth, 2016 to confirm the leave time and gauge to what extent understudies remain at the SRC by and large. Little's law [12] was then applied to assess the quantity of individuals in the SRC for a given time. Note that leave tallies were gathered uniquely for a solitary period and may contain predispositions (i.e., towards the finish of Spring semesters and toward the start of Summer semesters are normally substantially less swarmed than during Fall or Spring semesters, and the normal length of stay may be longer), yet we expect the normal span remains generally steady after some time with the goal that we can make predictions on the relative visit volumes to the SRC. Our optional dataset, got from WSU's Office of Institutional Research (IR), is segment data for clients who have visited the SRC between August seventeenth, 2012 to August seventeenth, 2015. This data included sex (female or male) and class level (first year recruit, sophomore, junior, senior, or graduate). A semester name (Fall 2012, Spring 2013, Summer 2013, and so forth.) was related with each record, demonstrating when the data was recovered. This detail is significant in light of the fact that the class level can change after some time. For instance, a first year recruit took a crack at Fall 2012 may turn into a sophomore in Spring 2013, contingent upon the quantity of credits taken. Like the timestamp data, the ID number segment was incorporated which is indistinguishable from the IDs in the timestamp data—one ID consistently alludes to the equivalent individual1. Cleaning. We at that point additionally handled and cleaned the data. In the timestamp data, we saw that some card exercises happened outside SRC's activity time (e.g., a card swipe at 4:30 am the point at which the SRC opens at 5 am). This is brought about by, for instance, a staff part swiping-in for work or a remote command being sent to the door to bolt/open. This ought not be considered as a piece of typical exercise data. Consequently, we sifted through this commotion (wrong data) by physically setting the quantity of individuals to zero for all times the SRC was shut (i.e., weekdays from 12 PM to 5 am and ends of the week from 12 PM to 9 am; occasion/summer hours may differ). We likewise considered the conceivable commotion of staff swipe-in for work during hours when the SRC is open. In light of SRC administrators' understanding, these records ought to be little in number comparative with the whole dataset (under 1%) and won't altogether influence our discoveries. We additionally made a section in the timestamp data that we loaded up with semester marks dependent on the date, permitting us to analyze the semester between datasets.

TABLE 1
Example of the merged timestamp and demographic data table. Header in *italics* indicates the source of data. "N/A" in the last row shows the scenario that a student 015374 has no demographic data in Spring 2017 since it is out-range of data collected.



CougarCard Center		*Office of Institutional Research*		
ID	time	sex	class level	semester
117256	2012-08-17 05:32:18	male	freshman	Fall 2012
746193	2012-08-17 05:45:09	female	senior	Fall 2012
...	...	...	...	...
026488	2015-03-15 21:14:37	male	graduate	Spring 2015
...	...	...	...	...
015374	2017-01-31 20:28:35	N/A	N/A	Spring 2017

The timestamp and segment datasets were then converged by ID number and semester mark to frame a table that contained all the data. A case of our consolidated data is appeared in Table 1. Be that as it may, some ID numbers had no or deficient segment records after the union. One explanation is that the time scope of the gathered segment dataset (through August seventeenth, 2015) is shorter than the timestamp dataset (through 31st, 2017). For inadequate examples, we physically filled in missing sections by expecting that if no records were found for the current semester for an ID, it continues as before as the 1. We might want to make reference to that in the first segment data acquired from IR, the ID numbers were genuine understudy numbers and were not indistinguishable from that in the timestamp data. The staff of the CougCard Center and the SRC required an extraordinary exertion in consolidating and coordinating all records into one spot at that point handed to us in a perfect arrangement, as to not damage understudy secrecy. past complete records discovered (all data arranged by date). For instance, if an understudy has a total arrangement of (unmistakable) records in both Spring 2014 and Fall 2014, however not Summer 2014, the Summer 2014 data is loaded up with the data of Spring 2014, not Fall 2014. We managed missing data for 32,431 understudies, a sum of 1.8 million missing passages out of 3 million records.

III.PROPOSED SYSTEM:

Initially, the current model can be additionally improved with additional top to bottom of Student nature and conduct Finally Our field of intrigue will be created and dependent on that activity warnings and recommendations will be made to upgrade our future in a legitimate manner This System likewise utilized in the following of understudies criminal operations at the school grounds and they can take essential activities on the spot to make sure about his future. Our work is comparative in soul to a couple of existing applications that are based on information found by gathering and dissecting action (regularly optionally related) data. This application assembles constant office utilization data at the Orlando Disney World and assessments holding up time with the goal that visitors can design their visits all the more productively.

To achieve our subsequent objective—helping understudies in determining when might be the best time to visit the SRC—we figured this time series prediction issue as a relapse task (rather than order as was recently done in [1]), where the objective is to foresee the quantity of individuals visiting the SRC for a given time span. A few models were worked for foreseeing visit volumes dependent on chronicled SRC use data. This segment first presents the approach and consequences of the three models we created. At that point a conversation on which model works the best for our errand follows. We think about modeling comes nearer from two fields of study, insights, and machine learning, and look at their appropriateness for our concern. As far as measurable modeling, we previously made an occasional guileless (Snaive) model to fill in as a pattern. This model "innocently" estimates future visit volume as the last watched an incentive from a similar season. Furthermore, we utilized one of the benchmark time series models, an autoregressive coordinated moving normal (ARIMA) model, which accomplished preferred prediction results over the gauge model. On the machine learning modeling side, we fit a random woodland (RF) model to the data and found that the RF model beat both the benchmark and the ARIMA model in our trial. All models were fabricated utilizing accessible bundles in the R programming condition [13]. We assessed the models by means of k-crease cross-approval [14], with k = 4. Specifically, we split the data into four-folds, as imagined in Figure 9. I th overlap contained I long stretches of preparing data (green square) with the beginning date being August twentieth, 2012. Around four months of data that follows the end date of preparing data were utilized for testing (yellow square). For instance, overlap 1 contained one year of preparing data from August twentieth, 2012 to August nineteenth, 2013 and testing data from August twentieth, 2013 to December 31st, 2013; crease 2 contained two years of preparing data from August twentieth, 2012 to August nineteenth, 2014, and testing data from August twentieth, 2014 to December 31st, 2014; and so on5. There are forgotten about data (dim squares) in all folds. Note that infold 4 we bar January 2017 data so the length of testing data is steady with overlay 1 to crease 3 (i.e., the closure date is December 31st, 2016 rather than January 31st, 2017). We assessed each model utilizing three assessment measurements and report their normal (±standard deviation) and best execution generally speaking folds on preparing and testing data. Initially, we use Root Mean Squared Error (RMSE, Equation 1), a broadly utilized technique in estimating model precision

$$RMSE = \sqrt{mean\left(e_t^2\right)} \qquad (1)$$

where et = yt yˆt is the forecast error for a time series t that computes the differences between the correct value y and the predicted value yˆ. The second evaluation metric is the seasonal variant of Mean Absolute Scaled Error (MASE, Equation 2), introduced in [15] and [10]. A scaled error qt is first defined as:

$$q_t = \frac{e_t}{\frac{1}{n-m}\sum_{i=m+1}^{n}|y_i - y_{i-m}|}$$

where m is the seasonal period of a time series t. The error qt is scaled by the in-sample Mean Absolute Error (MAE) from a naive method (in our case, this is the seasonal naive model) for each observation yi, comparing to the value from a previous season yim. Then the MASE is simply:

$$MASE = mean\left(|q_t|\right) \qquad (2)$$

According to [15], MASE can serve as an alternative to percentage error methods such as Mean Absolute Percentage Error (MAPE) = 100et yt . MAPE becomes infinite or undefined if the dataset contains yt = 0, which is a characteristic in our dataset (see Section 2, where we manually set the number of people to zero for all SRC closing times). Therefore, we believe MASE is a better measurement than the commonly used MAPE method for our dataset. Lastly, we apply Relative RMSE (RelRMSE, Equation 3) to compare the amount of performance improvement across models. When RelRMSE < 1, the proposed model is better than the baseline; when RelRMSE > 1, the proposed model is worse than the baseline:

$$RelRMSE = \frac{RMSE}{RMSE_b} \qquad (3)$$

where RMSEb means the RMSE of a benchmark result. For our situation, it is the occasional innocent model. For all tests, we utilize the best RMSE accomplished by each model to contrast and that of the occasional guileless model to figure RelRMSE.

CONCLUSION AND FUTURE WORK

In this article, we investigated and pictured patterns of understudy exercise movement as far as time and socioeconomics. We examined and looked at the prescient precision of an occasional credulous model, an ARIMA model, and a random woods model, and found that the random woodland model accommodates our dataset the best. The visit volume at the SRC for a given time span was precisely anticipated by the random backwoods model. The conveyed site page has helped the SRC representatives as far as day by day activities, for example, staff planning. We trust this work will fill in as a contextual investigation delineating the whole data science process, giving valuable bits of knowledge on how data could be gathered, handled, investigated, and utilized to make a client confronting data item. There are a few fascinating headings with regards to which this work could be stretched out later on. To start with, the current model can be additionally improved with a more inside and out analysis to accomplish an increasingly vigorous outcome. For instance, despite the fact that we evaluated to what extent understudies remain at the SRC on normal through manual checking, the way that understudies don't have to utilize their card to depart stays an issue in the event that we wish to foresee all out action time per client. One answer for this issue is to set up a mandatory swipe-out framework at the leave door that requires card swipes when leaving, with the impetus of 1) recording singular exercise times consequently and 2) use "length of remain" as an element to improve prediction exactness. Second, while the segment dataset is wealthy in data, we didn't fuse its explanatory outcomes as highlights in our present modeling strategies we might utilize the segment data to customize predictions and proposals for every client. Third, an increasingly broad philosophy can be detailed and tried. For instance, profound neural system models have increased extraordinary accomplishment in time series prediction issues throughout the years [26]. Since our timestamp dataset presents complex qualities, utilizing a profound model, for example, an intermittent neural system might accomplish better prediction execution.

REFERENCES

[1] Y. Du and M. E. Taylor, “Work In-progress: Mining the Student Data for Fitness,” in Proceedings of the 12th International Workshop on Agents and Data Mining Interaction (ADMI) (at AAMAS), Singapore, May 2016.

[2] InsiderGuide Inc., “Disney World Wait Times, Touring Plans Free by Undercover Tourist,” 2015, Accessed Jan 30, 2016. [Online]. Available: https://www.undercovertourist.com/apps/

[3] Waze Inc., “Waze - GPS, Maps and Social Traffic,” 2016, Accessed Jan 30, 2016. [Online]. Available: https://www.waze.com

[4] Google Maps, “Popular times,” 2016, Accessed Jan 30, 2016. [Online]. Available: https://support.google.com/business/ answer/6263531?hl=en

[5] E. Davami and G. Sukthankar, “Improving the performance of mobile phone crowdsourcing applications,” in Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2015, pp. 145–153.

[6] R. Schutt and C. O’Neil, Doing data science: Straight talk from the frontline. ” O’Reilly Media, Inc.”, 2013.

[7] R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O’Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, and F. Yasmeen, forecast: Forecasting functions for time series and linear models, 2018, r package version 8.4. [Online]. Available: http://pkg.robjhyndman.com/forecast

[8] R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. OTexts, 2014, ch. ARIMA models.

[9] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[10] R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. OTexts, 2014, ch. The forecasters toolbox.

[11] E. Alpaydin, Introduction to machine learning. MIT press, 2014.

[12] J. D. Little and S. C. Graves, “Little’s law,” in Building intuition. Springer, 2008, pp. 81–100.

[13] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[14] C. Bergmeir, R. J. Hyndman, B. Koo et al., “A note on the validity of cross-validation for evaluating time series prediction,” Monash University Department of Econometrics and Business Statistics Working Paper, vol. 10, p. 15, 2015.

[15] R. J. Hyndman, “Another look at forecast accuracy metrics for intermittent demand,” Foresight the International Journal of Applied Forecasting, pp. 43–46, 2006.

[16] R. D. C. Team, arima.c File Reference, 2016, R-devel Documentation 2016-04-14 SVN rev. 70486. [Online]. Available: http://docs. rexamine.com/R-devel/arima 8c source.html

[17] C. Chatfield, The analysis of time series: an introduction. CRC press, 2016.

[18] J. Geweke and S. Porter-Hudak, “The estimation and application of long memory time series models,” Journal of time series analysis, vol. 4, no. 4, pp. 221–238, 1983.

[19] A. M. De Livera, R. J. Hyndman, and R. D. Snyder, “Forecasting time series with complex seasonal patterns using exponential smoothing,” Journal of the American Statistical Association, vol. 106, no. 496, pp. 1513–1527, 2011.

[20] R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. OTexts, 2014, ch. Advanced forecasting methods.

[21] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, “Akaike information criterion statistics,” 1986.

[22] K.-S. Chan and B. Ripley, TSA: Time Series Analysis, 2012, r package version 1.01. [Online]. Available: https://CRAN.R-project.org/ package=TSA

[23] R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. OTexts, 2014, ch. Dynamic regression models.

[24] M. J. Kane, N. Price, M. Scotch, and P. Rabinowitz, “Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks,” BMC bioinformatics, vol. 15, no. 1, p. 276, 2014.

[25] A. Liaw and M. Wiener, “Classification and regression by randomforest,” R news, vol. 2, no. 3, pp. 18–22, 2002.

[26] M. Langkvist, L. Karlsson, and A. Loutfi, “A review of unsuper- ¨ vised feature learning and deep learning for time-series modeling,” Pattern Recognition Letters, vol. 42, pp. 11–24, 2014.