STORAGE GROWING FORECAST WITH BACULA BACKUP SOFTWARE CATALOG DATA MINING

Heitor Faria, Rommel Carvalho and Priscila Solis

Applied Computing Post Degree Program,
University of Brasilia (UnB), Brasilia, DF, Brazil

ABSTRACT

Backup software information is a potential source for data mining: not only the unstructured stored data from all other backed-up servers, but also backup jobs metadata, which is stored in a formerly known catalog database. Data mining this database, in special, could be used in order to improve backup quality, automation, reliability, predict bottlenecks, identify risks, failure trends, and provide specific needed report information that could not be fetched from closed format property stock property backup software database. Ignoring this data mining project might be costly, with lots of unnecessary human intervention, uncoordinated work and pitfalls, such as having backup service disruption, because of insufficient planning. The specific goal of this practical paper is using Knowledge Discovery in Database Time Series, Stochastic Models and R scripts in order to predict backup storage data growth. This project could not be done with traditional closed format proprietary solutions, since it is generally impossible to read their database data from third party software because of vendor lock-in deliberate overshadow. Nevertheless, it is very feasible with Bacula: the current third most popular backup software worldwide, and open source. This paper is focused on the backup storage demand prediction problem, using the most popular prediction algorithms. Among them, Holt-Winters Model had the highest success rate for the tested data sets.

KEYWORDS

Backup, Catalog, Data Mining, Forecast, R, Storage, Prediction, ARIMA, Holt-Winters

1. INTRODUCTION

By definition, backup data is only accessed in case of disaster [1], that is supposed to happen rarely. The first data scientist instinct would be to use this information also as a source for analytic engines, instead of fetching it from original source, without the concurrency with regular corporate workload as suggested by Poelker [2].

However, there is still another backup software information that is overlooked by the authors that has a lot of potential and is the scope of this practical work: the catalog database. It contains, for instance, the file locations from every backed-up platform, the duplicated files list, backup and restore jobs history etc.

Dhinaharan Nagamalai et al. (Eds) : CoSIT, SIGL, AIAPP, CYBI, CRIS, SEC, DMA - 2017
pp. 185– 196, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70418

The catalog learning can be also used to improve backup service itself, in order to identify error trends and capacity problems. A common challenge today, for example, is primary backup data continues to grow, more systems and data are deemed worth protecting, and backup retention is sometimes elongated as addressed by a recent Gartner Consultancy Report[3]. Digital data has snowballed to a level that frequently leads to backup storage capacity depletion, and it's imperative to predict these bottlenecks timely in order to avoid compromising backup data.

The purpose of the present work is to manifest that ARIMA and Holt-Winters forecasting models can provide the backup storage demand prediction, according to the terminated past backup jobs. In Section 2, we present the State-of-the-Art.

In the Section 3, we present the Related Work. The Section 4 shows the Methodology. In Section 5, we present the Results. Finally, the Section 6 draws some conclusions and final remarks. And Section 7, indicates Future Works.

March 01, 2017

2. STATE-OF-THE-ART

Bacula[1] is an open source backup software[4] whose metadata is stored in a database, e.g.: job logs, termination status, list of copied files with paths, storage media association, etc. According to Sibbald [5], it was the first published backup software to use a Structured Query Language and supports both MySQL[2] and PostgreSQL[3] open database services.

The Database Tables section of the Community Version manual [6] provides its full definition (table names, data types etc.), which is going to be the main researched data set of this work, only possible because of its open format.

According to Box et al. [7], a model that describes the probability structure of a sequence of observations is called a stochastic process that is a time series of successive observations is used in order to forecast the probably of distributions of future ones.

Time series are data series listed in time order [8], usually spaced with the same time frame and represented graphically through line charts. They can be also understood as streaming data with discrete values, and they have many tradition applications: mathematical finance [9], weather forecasting [10], intelligent transport [11] and econometrics [12]. Modernly, the DataMarket project [13] hosts varied time series data such as Crime, Ecology and Agriculture.

Box et al. [7] still describes Autoregressive Integrated Moving-Average (ARIMA) as a process more suitable to non-stationary time series observations (v.g.: stock prices) instead of auto-regressive (AR), moving average (MA) and mixed autoregressivemoving average (ARMA). Sato (2013), Pati and Shukla (2014), Wang et al. (2015), wrote recent papers using ARIMA, appearing as a relevant forecasting technique.

---

[1] http://blog.bacula.org/
[2] https://www.mysql.com
[3] https://www.postgresql.org/

Conforming to Goodwin and others [17], Holt-Winters is an exponential based method developed by C.C. Holt (1957) and P. Winters in (1960). As reported by Kalekar [18], it is used when the data exhibits both trend and seasonality (which are elements likely existent in this project observations). In line with Rodriguez et al. [19], exponential smoothing methods are based on the weighted averages of past observations, with the weights decaying exponentially as the observations get older. Puthran et al. (2014), Dantas et al. (2017), Athanasopoulos et al. (2017) and many other modern forecasting projects rely on Holt-Winters technique.

3. RELATED WORK

Until 2007, relevant backup book authors such as B Little and A. Chapa (2003) and Preston (2007) did not address the possibility of doing data mining in their studied backup software. Probably, they were moved because of the fact their studies mainly focused in proprietary backup software, where their databases have an unknown and closed format that is impossible to be read with third party software.

Guise (2008) was probably the first to write that backup software not only should, but must allow data mining of its metadata (among others): “without these features, a backup product is stagnant and not able to grow within an environment”. For the author, backup software should constitute value frameworks, but never monoliths.

Still, the impossibility of any developer to predict every possible data combination for backup report that a given company needs is also highlighted by Guise: *...the more useful feature of backup software for long-term integration into an enterprise environment is the ability to perform “data mining” - i.e., retrieving from the backup software and its database(s) details of the backups that have been performed.* More recently and even devoid of any scientific method, Poelker (2013) addressed the problem once more, suggesting and enumerating examples of how different types of backup data could be used for Knowledge Discovery and Data Mining, in order to aggregate value to that service.

That said, this work seems to be the very first in order to build data mining models for backup software databases and hopefully the foundation to other further analysis.

4. METHODOLOGY

According to Carvalho et al. (2014), CRISP-DM stands for Cross Industry Standard Process for Data Mining, which consists of a consortium oriented to establish an independent tool and industrial data mining process model. It aims to be generic enough to be deployed in different industry sectors, and sufficiently detailed to contemplate the conceptual phase of data mining project until the results delivery and validation.

Still, according to Carvalho et al., one of the methodology goals is to make data mining projects of any size run more efficiently: in a smaller time frame, more manageable, more inexpensive but the most fruitful.

The proposed life cycle of the data mining is presented in Figure 1 [26]. Data mining life cycle would consist of six phases [27], and their flow is not fixed: moving back and forth between different phases is usually required.

There is also an outer circle that represents the perpetual continuity of data mining process, since information and lessons fetch from a given project will very likely be use in further ones.And the inner arrows suggests the most frequent path between phases.

This will be the methodology used in this project, and the results will follow CRISP-DM phases.

Figure 1. Phases of the CRISP-DM Process Model

There is also an outer circle that represents the perpetual continuity of data mining process, since information and lessons fetch from a given project will very likely be use in further ones. And the inner arrows suggest the most frequent path between phases. This will be the methodology used in this project, and the results will follow CRISP-DM phases.

5. RESULTS

The results are presented ahead as the sections that represent the CRISP-DM executed steps for this work.

5.1. Business Understanding

Acquiring backup demanded storage is not an intuitive task. A naive approach would simply measure storage usage every given time in order to predict its growth. This would ignore already known information about future stored backup behavior, which are the retention times. Retention is the time frame within backup data cannot normally be discarded by the backup software, unless there is an exceptional human intervention. This has a significant impact in storage demand growing prediction, since monthly backup with one year of retention will demand twelve times more data storage occupation than retaining a backup for a single month, for example. A better way to predict backup storage growth is the cumulative sum of all terminated backup jobs during a time frame, subtracted by their own size after their expiration date (when their data is already disposable). For example, if in January 1st a 10GB backup is written, this amount is added to the demanded storage space total. It this job has 1 month retention, the same 10GB value is subtracted in February 1st. The goal is to use already known information to diminish prediction

algorithm margin error, since natural corporate backup size fluctuation demand can be very volatile by itself.

5.2. Data Understanding

Bacula MySQL database, in this case, is accessed using R database interface[4] and MySQL specific driver[5]. First a test database is used for model developing and testing, then validated with a production database.

Job and Pool tables are exported to a compatible R format. Job table contains the list of terminated backup Jobs information, including their size. Pool[6] table provides the jobs data retention times.

The job sizes (JobBytes) are expressed in bytes. Null values are discarded and the others converted to Gigabytes, since it is better for human reading. Decimals are rounded with digits, in order to not affect the total cumulative sum.

Backup jobs ending times used to build the time series are expressed originally with date, hours, minutes and seconds (YYYY-MM-DD hh:mm:ss). In order to simplify calculations and because it is insignificant for long term prediction, hour specification was trimmed and only dates are considered.

Retention times in the Pool table (VolRetention) is expressed in seconds. Those are rounded to days, because more significant and usual. Tables are merged so each Job now have their individual retention as a variable value.

Data frame is sort chronologically according to backup termination date, variables that supposed to be known as dates by R are set this way. Job sizes are sum in a cumulative function and a final storage size (sto.size variable) is calculated after the subtraction of already expired backup jobs. Data frame is padded with empty values when there is no backup jobs terminated on those days (necessary for time series). Also, jobs cumulative sum in those days receive last filled row value. There are 95 unique days with backups in this base for further validation reference.

5.3. Modeling

Time series (TS[7]) building R script runs dynamically, fetching first backup week and day from the current data set, in order to be able to work with any other Bacula database and different time frames.

---

[4] R DBI Library: https://cran.r-project.org/web/packages/DBI/
[5] RMySQL R Library: https://cran.r-project.org/web/packages/RmySQL/
[6] Pool is the group of backup volumes, or storage units, that has the same attributes such as retention.
[7] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ts.html

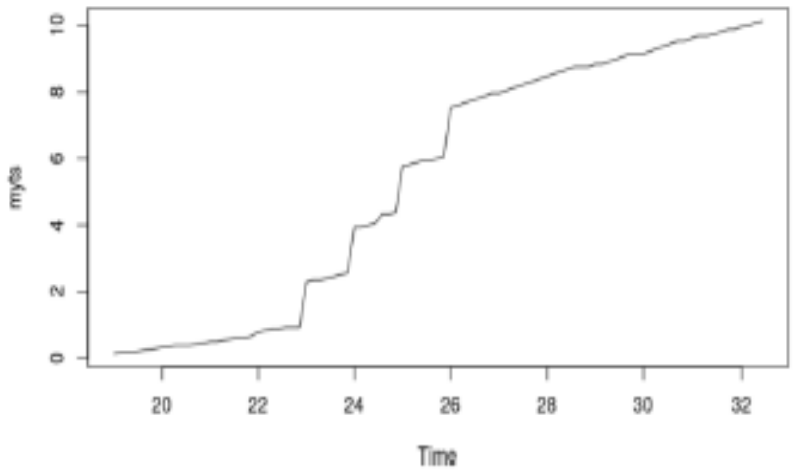


Figure 2. Test database storage size time series: total GB per week.

Figure 2 is a graphical representation of the created time series, and this will be used in order to develop the best models for storage size necessity prediction. It shows a higher initial growth ratio of backup storage size that corresponds to the very beginning of the backup system usage, and happens until backups fill their retention times and start to be recycled (discarded). This will probably be present in lots of Bacula production environments and affects the growing prediction in an undesired way. Since it is a very unpredictable behavior and backup configuration dependent, this is not filtered at this moment. The time scale is expressed in weeks, that is sufficient to run multiple backup jobs but not large enough to ignore huge backup size differences that may happen during a greater period. Last measured storage demand value is 10.1405GB.

5.3.1. ARIMA

In order to build the time series, 180 daily observations were provided, what would give a significant time frame of 6 months of predicted values.

The light gray areas of next plots represents the 80% prediction intervals while dark gray the 95% ones. The blue line is the predicted medium values. Future observation values near the blue line represents higher forecast accuracy.

The forecast library[8] provides the lower and upper prediction intervals (in this case, with 80% and 95% confidence), which are an estimated range of likely future values.

Figure 3 shows the model seems to be more affected by the initial state backup software operations (first 28 weeks) when the total size of backups grows faster since there are no prior terminated jobs. This can lead to misleading results in fresh production environments and long term predictions. The last forecast value after 6 months of forecast is 25.97GB.

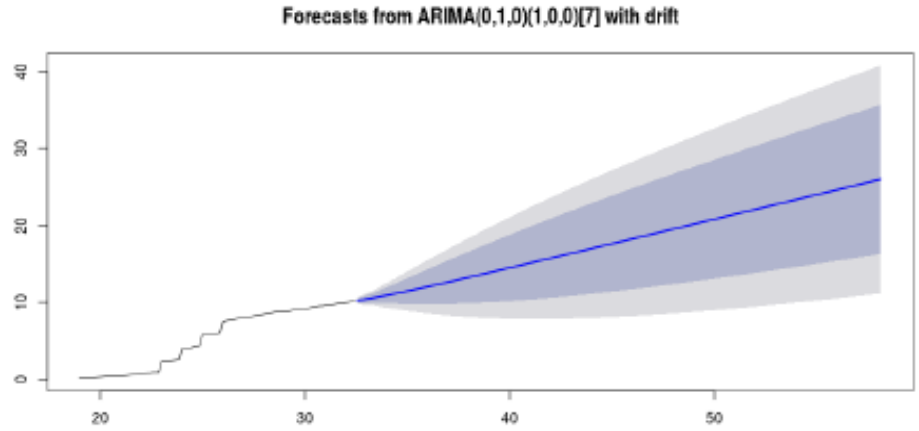


Figure 3. ARIMA model (total GB per weeks).

5.3.1. Holt-Winters

Holt-Winters[9] exponentially weighted moving averages is more tolerant to the misleading first 28 weeks of the initial state of the backup software, which would provide more reliable prediction values as shown in Figure 4. The forecast value for backup storage for a 6 months time frame is 20.77GB

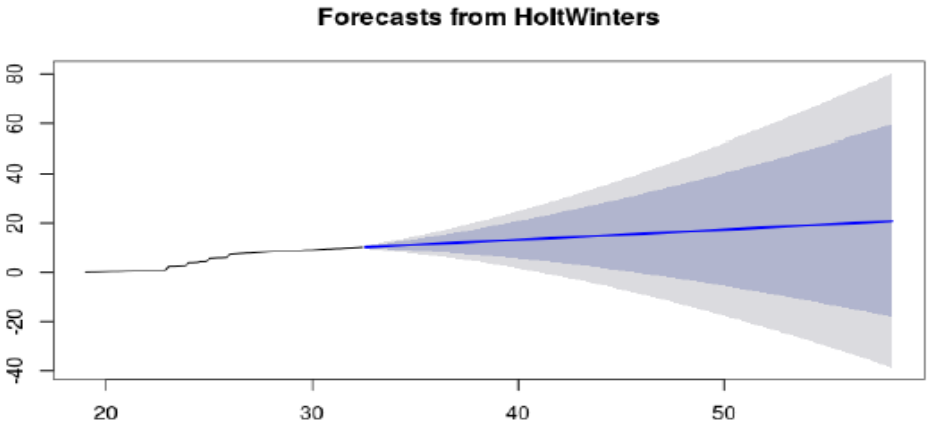


Figure 4. Holtwinters 6 months forecast (total GB per weeks)

5.4. Evaluation

Holt-Winters model responds quicker to trend changes such as lower storage growing trend after initial backup system deploy, being considered more suitable to cope with production environments data. It is the chosen model for testing and validation.

Figure 5 presents, in the same scale, the same used Holt-Winters prediction model against an approximately 30% larger dataset filled with real values. Last forecast storage size is 12.20GB, 10.61% higher than the the actual cumulative backup jobs sum is 11.03GB, but still in the 80% prediction interval.

Another random Bacula production environment database is used in order to apply the selected model. Forecast section corresponds to a slice of approximately 60% entries of the data frame.

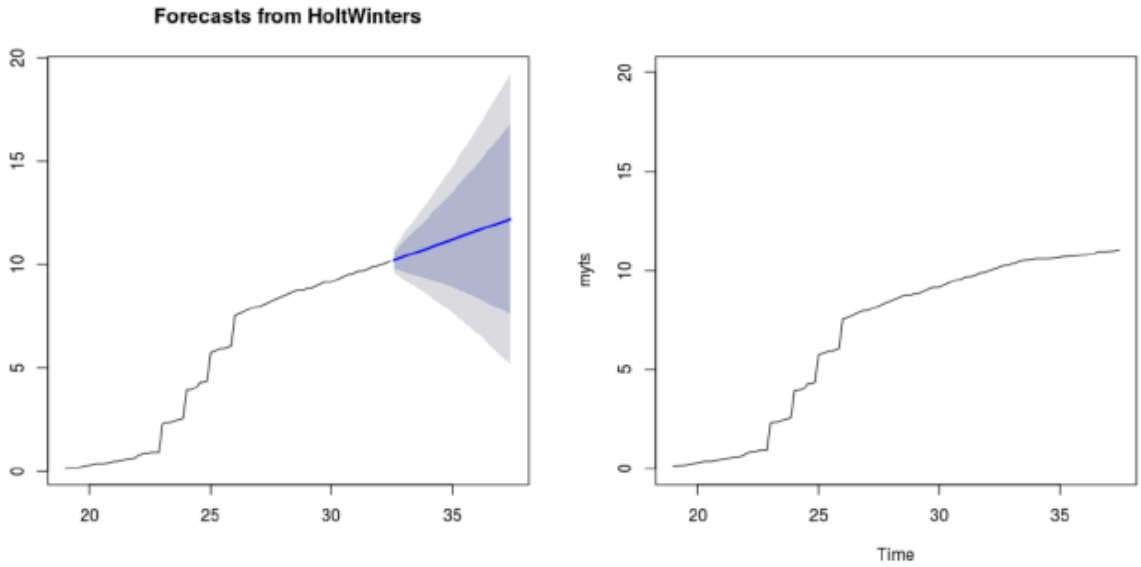


Figure 5. Holt-Winters forecast against test data set storage growing (total GB per weeks).

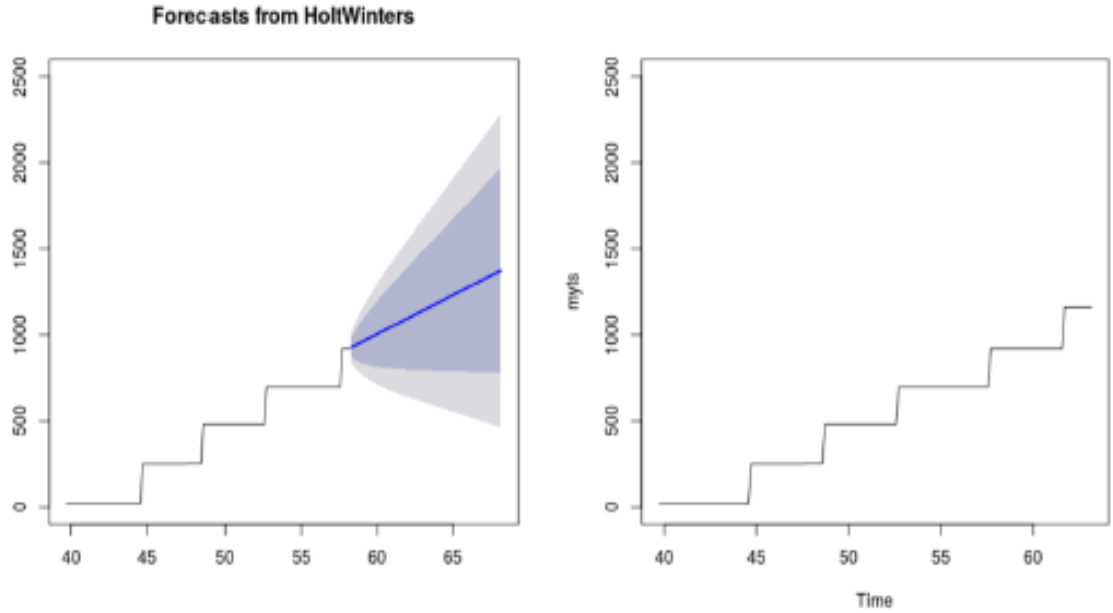


Figure 6. Holt-Winters forecast against production actual data (total GB per weeks).

As displayed in Figure 6, the forecast value (left plot) for ten weeks of prediction (1374.19GB) is 18.33% greater than the actual (right plot) storage size (1161.35GB). However, the lower 95% prediction confidence interval limit for the forecast model is 778.23GB (dark gray area), so the real storage forecast size is into it. The model is satisfactory for available datasets.

6. CONCLUSION

Hyndman [28] study found prediction intervals calculated to include the true results 95% of the time only get it right between 71% and 87% of the time.

In this way, the formula to calculate backup storage cumulative sum for storage size and the choice of the Holt-Winters Model is suitable for the current type of data and for a reasonable and specially vegetative backup size growth, being able to forecast storage growth for a significant amount of time within the 95% prediction interval.

As a remark, it is known that IT infrastructure teams needs to cope with series of unplanned changes from different corporate areas, and for those there are currently no models that could handle them.

The chosen Holt-Winters model must be applied to other production sets of information of different sizes, in order to be considered successfully deployed, which would be the last CRISPDM stage.

Another backup database data mining project execution might also produce the results bellow, among others:

- Predict backup clients demand for restore;
- Fetch all time typical and current successful terminated backup jobs streak;
- Classify backup clients performance (clusterization);
- Identify current backup window compliance;
- Match hardware set suggestions in order to attend current performance and storage demand;
- Analyze the potential benefit of using a file level deduplication feature;
- Analyze the potential benefit of using a block global deduplication feature;
- Identify jobs which behave unexpected with execution log information mining;
- Suggest disk based backup ideal backup volume size.

ACKNOWLEDGEMENTS

Heitor Faria would like to thank Roberto Mourao for the help with the R mechanics and Wanderlei Huttel for providing the production database analyzed in this paper.

REFERENCES

[1] W. C. Preston, Backup and Recovery, 2007.

[2] C. Poelker, "Backups as a source for data mining," 2013. [Online]. Available: http://www.computerworld.com/article/2474646/data-center/backups-as-a-source-for-datamining.html

[3] Pushan Rinnen and D. Russel, "Challenging Common Practices for Backup Retention," Gartner, Inc., USA, Tech. Rep. G00278794, Jul. 2015.

[4] X. Zhang, Z. Tan, and S. Fan, "NSBS: Design of a Network Storage Backup System," 2015.

[5] K. Sibbald, "Bacula Problem Resolution Guide," Aug. 2015. [Online]. Available: http://www.bacula.org/7.2.x-manuals/en/ problems/Problem Resolution Guide.html

[6] ——, "Database Tables. Bacula Developer's Guide," Aug. 2011. [Online]. Available: http://www.bacula.org/5.1.x-manuals/en/ developers/developers/Database Tables.html

[7] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control. John Wiley & Sons, Jun. 2015, googleBooks-ID: lCy9BgAAQBAJ.

[8] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," in Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, ser. DMKD '03. New York, NY, USA: ACM, 2003, pp. 2–11. [Online]. Available: http://doi.acm.org/10.1145/882082.882086

[9] M. Corazza and C. Pizzi, "Mathematical and Statistical Methods for Actuarial Sciences and Finance || A skewed GARCH-type model for multivariate financial time series," 2010. [Online]. Available: http://booksc.org/book/21982468

[10] R. Honda, S. Wang, T. Kikuchi, and O. Konishi, "Mining of Moving Objects from Time-Series Images and its Application to Satellite Weather Imagery," Journal of Intelligent Information Systems, vol. 19, no. 1, pp. 79–93, 2002. [Online]. Available: http://link.springer.com/article/10.1023/A:1015516504614

[11] L. A. James, "Sustained Storage and Transport of Hydraulic Gold Mining Sediment in the Bear River, California," Annals of the Association of American Geographers, vol. 79, no. 4, pp. 570–592, Dec. 1989. [Online]. Available: http://onlinelibrary.wiley.com/doi/ 10.1111/j.1467-8306.1989.tb00277.x/abstract

[12] J. E. H. Davidson, D. F. Hendry, F. Srba, and S. Yeo, "Econometric Modelling of the Aggregate TimeSeries Relationship Between Consumers' Expenditure and Income in the United Kingdom," The Economic Journal, vol. 88, no. 352, pp. 661–692, 1978. [Online]. Available: http://www.jstor.org/stable/2231972

[13] "Time Series Data Library - Data provider," 2017. [Online]. Available: https://datamarket.com/

[14] R. C. Sato, "Gerenciamento de doenas utilizando sries temporais com o modelo ARIMA,"Einstein (So Paulo), 2013. [Online]. Available: http://www.repositorio.unifesp.br/handle/11600/7670

[15] J. Pati and K. K. Shukla, "A comparison of ARIMA, neural network and a hybrid technique for Debian bug number prediction," in Computer and Communication Technology (ICCCT), 2014 International Conference on. IEEE, 2014, pp. 47–53. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7001468/

[16] W.-c. Wang, K.-w. Chau, D.-m. Xu, and X.-Y. Chen, "Improving Forecasting Accuracy of Annual Runoff Time Series Using ARIMA Based on EEMD Decomposition," Water Resources Management, vol. 29, no. 8, pp. 2655–2675, Jun. 2015. [Online]. Available: http://link.springer.com/10.1007/s11269-015-0962-6

[17] P. Goodwin and others, "The holt-winters approach to exponential smoothing: 50 years old and going strong," Foresight, vol. 19, pp. 30–33, 2010. [Online]. Available: https://www.researchgate.net/profile/Paul Goodwin/ publication/227439091 The Holt-Winters Approach to Exponential Smoothing 50 Years Old and Going Strong/links/0046351dc5a91a08de000000.pdf

[18] P. S. Kalekar, "Time series forecasting using holtwinters exponential smoothing," Kanwal Rekhi School of Information Technology, vol. 4329008, pp. 1– 13, 2004. [Online]. Available: https://c.forex-tsd.com/ forum/69/exponentialsmoothing.pdf

[19] H. Rodriguez, V. Puig, J. J. Flores, and R. Lopez, "Combined holt-winters and GA trained ANN approach for sensor validation and reconstruction: Application to water demand flowmeters," in Control and Fault-Tolerant Systems (SysTol), 2016 3rd Conference on. IEEE, 2016, pp. 202–207. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7739751/

[20] D. Puthran, H. C. Shivaprasad, K. K. Kumar, and M. Manjunath, "Comparing SARIMA and HoltWinters forecasting accuracy with respect to Indian motorcycle industry," Transactions on Engineering and Sciences, vol. 2, no. 5, pp. 25–28, 2014. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.437.1043&rep=rep1&type=pdf

[21] T. M. Dantas, F. L. Cyrino Oliveira, and H. M. Varela Repolho, "Air transportation demand forecast through Bagging Holt Winters methods," Journal of Air Transport Management, vol. 59, pp. 116–123, Mar. 2017. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0969699716302265

[22] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and F. Petropoulos, "Forecasting with temporal hierarchies," 2015. [Online]. Available: https://mpra. ub.unimuenchen. de/id/eprint/66362

[23] D. B Little and D. A. Chapa, Implementing Backup an Recovery: The Readiness Guide for the Enterprise, 2003.

[24] P. d. Guise, Enterprise Systems Backup and Recovery: A Corporate Insurance Policy. CRC Press, Oct. 2008, google-Books-ID: 2OtqvySBTu4C.

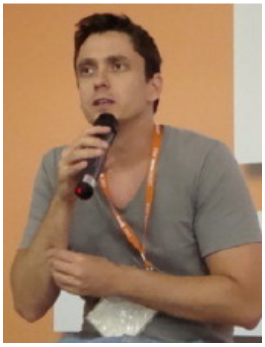
[25] R. N. Carvalho, L. J. Sales, H. A. D. Rocha, and G. L. Mendes, "Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil," 2014. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=77038E9B372F7790F8F0FFDE0A3BF3C1? doi=10.1. 1.662.1132

[26] K. Jensen, "English: A diagram showing the relationship between the different phases of CRISPDM and illustrates the recursive nature of a data mining project." Apr. 2012. [Online]. Available: https://commons.wikimedia.org/wiki/File: CRISP-DM Process Diagram.png

[27] R. Wirth, "CRISP-DM: Towards a standard process model for data mining," 2000, pp. 29–39.

[28] R. Hyndman, “A state space framework for automatic forecasting using exponential smoothing methods,” Jul. 2002. [Online]. Available: http://robjhyndman. com/papers/hksg/

AUTHORS



Heitor Faria was entitled with the Alien of extraordinary ability visa by the US Government for his work in Information Technology and Open Source Software. Master degree at Brasília National University (UNB). "Bacula: Open Source Backup Software" (English & Portuguese) and "Open Licenses & Fundamental Rights" books author (Poruguese). Bacula, Alfresco and Wordpress Training instructor at Udemy, with more than 800 students in 46 countries. Works as a System Analyst on a brazilian governmental company called SERPRO and for Neocode Software (Canada). Law Graduated. IT Service Manager and Project Management extension degrees. Bacula brazilian community founder. Has plenty of experience as server/backup systems administrator (Windows, Linux, Netware, directory services) and as IT / Project manager. Speaker at several international open source software events. ITIL-F, TOEFL and LPIC-III certificated professional.

Rommel Novaes Carvalho is a researcher with the Brazilian Office of the Comptroller General and an affiliate professor at the University of Brasília. His research focus is on uncertainty in the Semantic Web using Bayesian Inference, Data Science, Software Engineering, as well as Java, R, and Python programming. He is the developer of PR-OWL (v2.0) and UnBBayes, an open source, java-based graphical editor for Multi-Entity Bayesian Network and Probabilistic Web Ontology Language (PROWL), among other probabilistic graphical models.

Priscila Solis Barreto is professsor at the Computer Science Department, University of Brasília. From May/2002 until March/2007, she was a doctoral candidate at the University of Brasília, Department of Electrical Engineering. Her doctoral thesis subject was traffic characterization in multimedia networks. In 2000, she finished her master studies at the Federal University of Goiás, School of Electrical Engineering. Her master thesis subject was in traffic prediction for ATM networks. Priscila worked for several years as a Computer Science Technician and Consultant, developing projects and software for different entities, such as Terra Networks, DR Sistemas and CIAT-BID. She was also a professor at the Catholic University of Goiás for several years. After her undergraduate studies, Priscila had a scholarship to work as a researcher, for one year, in the DIST (Dipartamento de Informática, Sistemística e Telemática), at the University of Genova, Italy.

Her undergraduate studies were made at the Faculty of Engineering in Computer Science, University Francisco Marroquin, in Guatemala City, Guatemala. Currently, Priscila continues her research at the COMNET Lab within the LARF Research Group. Her main research interests are multimedia traffic models, network performance, network planning and traffic engineering. Priscila is also interested in the development of software tools for simulation. Another recent topic she’d began to work is Future Internet and Trust and Security in Wireless Networks.