The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2022

Volume 21, Pages 39-45

IConTES 2022: International Conference on Technology, Engineering and Science

Wireless Channel Availability Forecasting with a Sparse Geolocation Spectrum Database by Penalty-Regularization Logistic Models

Vladimir II Christian OCAMPO
De La Salle University

Lawrence MATERUM
De La Salle University

Abstract: Television uses electromagnetic waves that carry audio and video. The unused frequencies or channels in broadcasting services are referred to as television white spaces. The unused spectrum can be managed to provide internet access in coordination with surrounding TV channels to avoid interference. Different ways of dynamically managing spectrum management have been conceived, and geolocation databases are considered the better option. Geolocation databases, when updated and complete, are helpful when frequencies are dynamically shared. In real life, the spectrum availability for a secondary user lacks numerous information; hence, it is sparse. This paper forecasts wireless channel availability given a sparse geolocation spectrum database. A dynamic sparse forecasting model is proposed through logistic penalized regression. Results show that forecasting accuracy is mostly above 90% on average when sparsity penalty terms are incorporated into the model. Forecasting accuracy is improved when penalty terms are integrated into the logistic regression models to account for sparsity.

Keywords: Channel availability, Forecasting, Geolocation database, Penalized logistic regression, TVWS

Introduction

Television (TV) uses electromagnetic (EM) waves that carry modulated audio and video to broadcast. The TV EM waves have long wavelengths and travel far distances. The spectrum is divided into channels, which may be occupied or not. An unused portion of the spectrum is called white space.

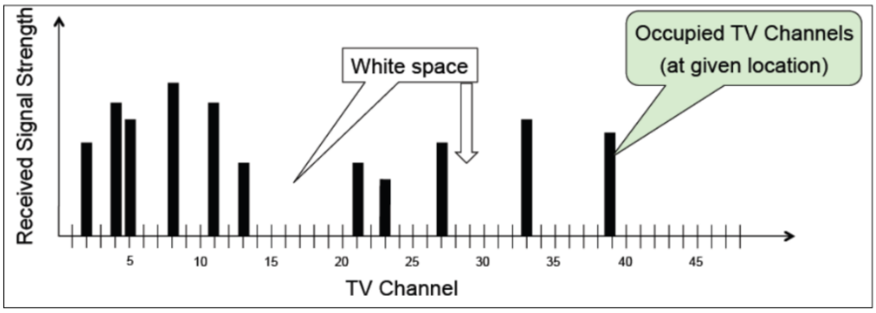

Figure 1. Concept of white space.

TV White Space (TVWS) communications involves using the white spaces in the TV spectrum for other non-broadcasting purposes such as point-to-point connectivity, Internet services, emergency communications, and the like. Studies have explored using these white space bands with an increased demand for network usage and remote areas needing communication access. Some applications on TVWS include environment monitoring,

- This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 Unported License, permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

- Selection and peer-review under responsibility of the Organizing Committee of the Conference

© 2022 Published by ISRES Publishing: www.isres.org

emergency tracking, and communication that can reach far-flung and remote areas, which are hard to reach using traditional communication technologies, such as the Internet. TVWS communication is a growing technology allowing dynamic frequency spectrum usage, which can benefit rural and underserved areas of developing countries like the Philippines. According to Inquirer, a map of digital poverty in the Philippines discovered that rural areas have no access to sufficient Internet speeds. This outcome means that rural areas are left behind in terms of connectivity (Sy, 2021). TVWS can be used for a super-high-speed Wi-Fi, called Wi-Fi 2.0, that can provide access to underserved areas. This Wi-Fi can reach gigabits per second speeds (Sarkar et al., 2016). TVWS has been tried in different countries, such as Japan, India, Singapore, and even the Philippines (Mody, 2017).

There is a drawback to TVWS, at least when using existing technologies. TVWS uses dynamic sharing of frequencies across space and time. In addition, TVWS has no protection from interference, especially from primary incumbents like TV stations. So, there must be a way to coordinate the white space devices (WSD). Spectrum sensing was one way of coordinating WSDs where the device senses the unoccupied frequencies. However, design and performance were challenging, as tested by the Federal Communications Committee (FCC) (Gurney et al., 2008). Geolocation databases are better options because they contain information such as occupied frequencies at a given location and time. A WSD can query the geolocation database, and the geolocation database responds with the information needed by the WSD, such as frequency availability. However, geolocation databases are sparse, and no study has attempted to forecast using sparse TVWS geolocation databases, giving this research the aim to address this problem.

Literature Review

Previous studies explored methods of spectrum management so that dynamic sharing of the spectrum between primary and secondary devices without interference with primary users is achieved. Martin et al. (2016) developed the General Enhanced Detection Algorithm (GEDA) to enhance the detection of primary users (PU) and optimize the secondary user's Quality-of-Service (QoS). Bourdena et al. (2012) developed a real-time secondary spectrum market (RTSSM) through a spectrum broker. Zhao et al. (2015) used game theory in dynamic spectrum management.

Some studies dealt with the TVWS geolocation database. Deep learning methods were applied to make predictions using a geolocation database in the study of Shawel et al. (2019). However, there are no studies yet involving sparse geolocation data.

Hurley & Rickard (2009) compared numerous sparsity measures. Benarabi et al. (2021) used density, which is the ratio of the number of non-zero elements to the total number of elements in a vector, as a basis for its sparsity measure. Goswami et al. (2018) measured the sparsity of a network graph using measures based on the Gini index. The integration of sparsity in predictive analytics has not yet been explored.

Some studies investigated forecasting using time series models. Wu et al. (2012) derived the Adversarial Sparse Transformer (AST) from Generative Adversarial Networks (GANs). Ardia et al. (2019) employed time series aggregation using computed textual sentiment to forecast high-dimensional data. Flaxman et al. (2019) propose a generic spatio-temporal forecasting method for a challenge in predicting crimes in real time. Spatio-temporal forecasting was also performed on sparse data to predict urban traffic flow (Zheng et al., 2020). The space and time complexity of models increase if data are sparse. Model algorithms might behave in extraordinary results if sparsity is present. No studies have considered the sparsity factor in forecasting models, especially spectrum forecasting. Geolocation prediction capabilities have been developed primarily using deep learning models. However, no studies have explored using sparse forecasting capabilities for geolocation databases. This area is a potential novel contribution of this study.

Sparsity

Sparsity can be thought of as having a small amount of information or packing a large proportion of energy in a small number of coefficients (Hurley & Rickard, 2009). For instance, a matrix with many zeros is referred to as sparse. The concept of sparsity is widely used in various areas of research. Areas such as ocean engineering, antennas and propagation, and image processing employ this concept (Hurley & Rickard, 2009). In addition, sparsity was a central concept for the success of machine learning algorithms and numerous techniques such as matrix factorization, signal processing, dictionary learning, and support vector machines. Because the concept of sparsity could be abstract, many measures of sparsity were introduced.

$$\|\mathbf{c}\|_p = \left(\sum_{j=1}^{N} |c_j|^p\right)^{\frac{1}{p}} \tag{1}$$

The majority of the studies use ℓ^p-based sparsity measures, with ℓ^0 and ℓ^1 being most common. Given a vector $\mathbf{c} \in \mathbb{R}^N$, the traditional definition of ℓ^p norm is described in (1). The ℓ^0 measure simply counts the number of non-zero elements in a vector. It is the traditional measure in a lot of mathematical settings. However, ℓ^0 is difficult to solve as the derivative of the measure has no information (Hurley & Rickard, 2009), and problems involving ℓ^0 are combinatorial in nature. Given this, the ℓ^1 measure is usually an approximation of the ℓ^0 norm and is being employed in many optimization problems. ℓ^1 can be used as a penalty. Other ℓ^p based measures use $0 < p < 1$, such as the study of Xu et al. (2010) that proposed the use of $\ell^{\frac{1}{2}}$ norm. Other sparsity measures are the Hoyer and Gini indices, as defined in (2) and (3), respectively.

$$\mathrm{HI} = \left(\sqrt{N} - \frac{\sum_j c_j}{\sqrt{\sum_j c_j^2}}\right)\left(\sqrt{N} - 1\right)^{-1} \tag{2}$$

$$\mathrm{GI} = 1 - 2\sum_{k=1}^{N} \frac{c_{(k)}}{\|\mathbf{c}\|_1}\left(\frac{N - k + \frac{1}{2}}{N}\right) \tag{3}$$

given ordered data: $c_{(1)} \leq c_{(2)} \leq \cdots \leq c_{(N)}$

Logistic Regression and Regularization

A logistic regression models variables whose response is categorical. A categorical variable takes on discrete values and does not use the ratio scale (Czepiel, 2002). Logistic regression derives from the generalized linear model developed by Nelder and Wedderburn. For a binary logistic model, the response variable has two possible values, and the model is defined in (4).

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{\mathrm{k}=0}^{K} x_{ik}\beta_k = (\mathbf{X}\boldsymbol{\beta})_i \tag{4}$$

In (4), π_i represents the probability of success, x_{ik} represents the predictor variables and β_k is a parameter of the model. The predictor variables can also be put in a matrix $\mathbf{X} \in \mathbb{R}^{N\times(K+1)}$, and $x_{i0} = 1$ for $i = 1,2, \ldots, N$. Unlike linear models, where parameters are found by minimizing the squares of the errors—the Least Squares Estimate—logistic models use maximum likelihood. For a binary logistic model, the likelihood is the binomial probability distribution.

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \binom{n_i}{y_i} \pi_i^{y_i}(1 - \pi_i)^{n_i - y_i} \tag{5}$$

When estimating the parameters by maximum likelihood, the log-likelihood is maximized, or equivalently, the negative of the log-likelihood is minimized, as in (6). However, Czepiel (2002) noted that estimating the maximum likelihood is computationally intractable, and numeric methods are employed instead.

$$\arg\min_{\boldsymbol{\beta}} - \ln L(\boldsymbol{\beta}|\mathbf{y}) \tag{6}$$

Binary logistic regression can be extended to form Multinomial Logistic Regression (MLR), in which the response variable can take at least two values. For an MLR model, given J discrete categories of the response where $J \geq 2$, with the Jth category as the baseline, the model is defined in (7). The probability distribution of the multinomial response variable y is defined in (8).

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \log\left(\frac{\pi_{ij}}{1-\sum_{j=1}^{J-1}\pi_{ij}}\right) = \sum_{k=0}^{K} x_{ik}\beta_{kj} \qquad (7)$$

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N}\left[\frac{n_i!}{\prod_{j=1}^{J} y_{ij}!}\cdot\prod_{j=1}^{J}\pi_{ij}^{y_{ij}}\right] \qquad (8)$$

Methodology

For predicting channel availability, penalized logistic regression models have been formulated. From the whole data set, 80% are relegated training, and 20% are for testing. The training and testing data points are randomly selected, and the testing data is used to evaluate the model. Online sources include primary user information such as company name, channel, and location (longitude and latitude). Channel availability is surveyed from 00:00H (12:00 AM) to 23:59H (11:59 PM) in 30-minute intervals. For a secondary user, the channel is deemed available if the primary user broadcasting on the channel is off-air or does not interfere with a primary user. Availability is labeled with a 1 for available channels and 0 for unavailable channels. The dataset is available online (Ocampo, Vladimir, n.d.). The model input-process-output diagram is illustrated in Figure 2.

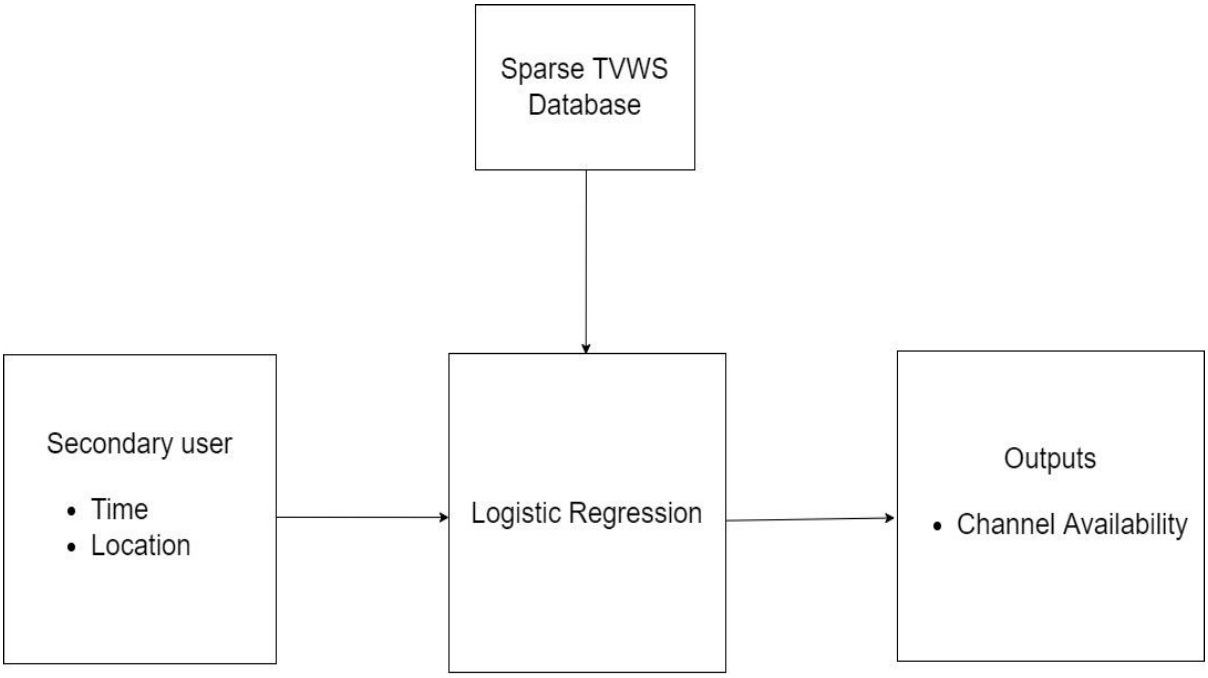

Figure 2. Input-process-output diagram

Penalized logistic models have been developed to predict channel availability. The training and test data points are randomly selected. The penalties included in the study are Lasso (ℓ^1), Ridge (ℓ^2), and Elastic net (ℓ^1 + ℓ^2). The analysis has been carried out through R with the source codes available online (Vladimir Ocampo, 2022).

For evaluating the model, accuracy is used. This metric is determined by hits, misses, false alarms, and correct rejections. A hit (true positive) is called when an available channel is predicted available. A miss (false negative) is called when an available channel is predicted as unavailable. A false alarm (false positive) is called when an unavailable channel is predicted as available. A correct rejection is called when an unavailable channel is predicted as unavailable. Table 1 summarizes the evaluation criteria of hits, misses, false alarms, and correct rejections.

Table 1. Forecast quality evaluation criteria

	Forecasted TVWS channel availability	
Observed TVWS channel availability	Available	Not Available
Available	Hit (True Positive)	Miss (False Negative)
Not Available	False Alarm (False Positive)	Correct Rejection (True Negative)

The accuracy is defined in (9).

$$\%\text{Accuracy} = \frac{\text{Hit} + \text{ Miss}}{\text{Total samples in test data}} \times 100\% \qquad (9)$$

Results and Discussion

In predicting the channel availability for TVWS, logistic regression with and without regularization is developed. The inputs include day and time and whether there is an interference or not. The model has been carried out for eight channels: 4, 5, 7, 9, 11, 13, 27, and 37. Thirty trials have been conducted, where in each trial, training and test data are randomly selected from the whole data set. The model is evaluated in terms of accuracy, precision, and recall, and the averages of those values are taken from those thirty trials. Figure *3* shows a bar graph of the average accuracies of logistic regression (Logit), cross-validated lasso (Cvlasso), cross-validated ridge (Cvridge), and cross-validated elastic net (Cvelnet).

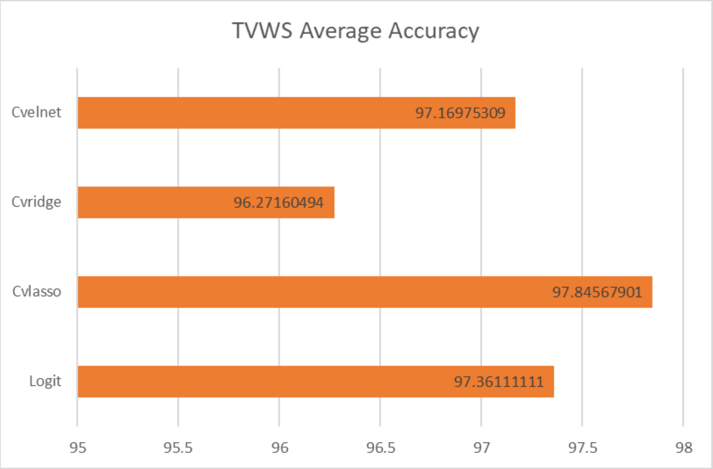

Figure 3. Average accuracies of TVWS logistic models.

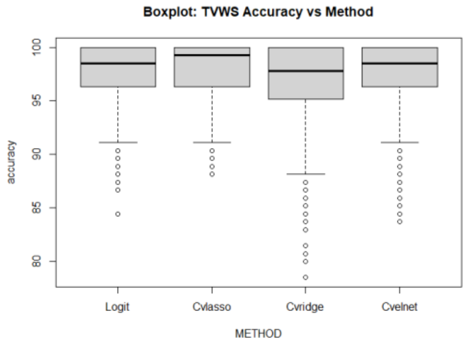

Figure 4. Boxplot of the accuracies of the model (generated with R)

In addition, a statistical analysis is carried out to determine if regularization has a significant effect. For each trial, the accuracy of the model is taken. Using the Shapiro-Wilk test, the accuracy data is not normally distributed; thus, Kruskal-Wallis was carried out to determine the significance of regularization. The boxplot is shown in Figure *4*. The Kruskal-Wallis test resulted in a significant difference among groups. However, the boxplot shows that regularized logistic regression performs just as well as unregularized logistic regression in terms of accuracy. Post-hoc tests confirm that was the case, seeing a difference between elastic net and ridge regressions, probably due to the presence of outliers. The insignificant difference between regularized and unregularized logistic models may be because the input variables are limited when there could have been more variables in reality.

Conclusion

Due to sparse spectrum availability for a secondary user, the paper formulated logistic regression models with sparsity regularization to forecast wireless coverage and frequency availability in sparse geolocation spectrum databases. The sparsity penalty terms incorporated in the logistic regression models are Lasso (ℓ^1), Ridge (ℓ^2), and Elastic Net ($\ell^1 + \ell^2$). The average accuracies of the models range from 96.27% to 97.85%. Statistical analyses indicate that regularized logistic regression has no significant difference from unregularized logistic regression in terms of accuracy, which could be due to a limited number of variables used in the model.

Recommendations

Given that the regularized logistic regression performs just as well as the unregularized logistic regression, the number of input variables may be limited. A future direction for this research would include more input variables. In addition, other sparsity penalty terms can be explored, such $\ell^{\frac{1}{2}}$ and a function of the Fisher information matrix.

Scientific Ethics Declaration

The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM journal belong to authors.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Technology, Engineering and Science (www.icontes.net) held in Antalya/Turkey, on November 16-19, 2022.

* The researcher would like to thank DOST ERDT for the funding and support for the research.

References

Ardia, D., Bluteau, K., & Boudt, K. (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting*, *35*(4), 1370–1386.

Benarabi, T., Adnane, M., & Mansour, M. (2021). Energy and sparse coding coefficients as sufficient measures for VEBs classification. *Biomedical Signal Processing and Control*, *67*, 102493. https://doi.org/10.1016/j.bspc.2021.102493

Bourdena, A., Mastorakis, G., Pallis, E., Arvanitis, A., & Kormentzas, G. (2012). A dynamic spectrum management framework for efficient TVWS exploitation. *2012 IEEE 17th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 51–55.

Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: Theory and implementation. *Available at Czep. Net/Stat/Mlelr. Pdf, 83*.

Flaxman, S., Chirico, M., Pereira, P., & Loeffler, C. (2019). Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ "Real-time crime forecasting challenge." *The Annals of Applied Statistics*, *13*(4), 2564–2585.

Goswami, S., Murthy, C., & Das, A. K. (2018). Sparsity measure of a network graph: Gini index. *Information Sciences*, *462*, 16–39.

Gurney, D., Buchwald, G., Ecklund, L., Kuffner, S. L., & Grosspietsch, J. (2008). Geo-location database techniques for incumbent protection in the TV white space. *2008 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, 1–9.

Hurley, N., & Rickard, S. (2009). Comparing measures of sparsity. *IEEE Transactions on Information Theory*, *55*(10), 4723–4741.

Martin, J. H., Dooley, L. S., & Wong, K. C. P. (2016). New dynamic spectrum access algorithm for TV white space cognitive radio networks. *IET Communications*, *10*(18), 2591–2597.

Mody, A. (2017). Tutorial on whitespaces, technologies and standardization... Means to bridge the digital divide. *IEEE P802 22-17-0054-Rev0/Ec-17-0147-00-WCSG*, 1–81.

Ocampo, V. (n.d.). *A 2022 television white space geolocation database of the greater manila area of the Philippines*. IEEE DataPort. https://doi.org/10.21227/YZ26-E616

Sarkar, B. D., Shankar, S., Verma, S., & Singh, A. K. (2016). Utilization of television white space for high speed Wi-Fi application TVWS usage. *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, 240–243.

Shawel, B. S., Woldegebreal, D. H., & Pollin, S. (2019). Convolutional LSTM-based long-term spectrum prediction for dynamic spectrum access. *2019 27th European Signal Processing Conference (EUSIPCO)*, 1–5.

Sy, S., Araneta, A., Rahemtulla, H., Carrasco, B., & Balgos, S. (2021). *Mapping digital poverty in PH*. INQUIRER.Net. https://business.inquirer.net/318223/mapping-digital-poverty-in-ph

Vladimir, O.. (2022, October 27). *Wireless channel availability forecasting with a sparse geolocation spectrum database by penalty-regularization logistic models*. Code Ocean. https://codeocean.com/capsule/9352952/tree

Wu, Q., Law, R., & Xu, X. (2012). A sparse Gaussian process regression model for tourism demand forecasting in Hong Kong. *Expert Systems with Applications*, *39*(5), 4769–4774.

Xu, Z., Zhang, H., Wang, Y., Chang, X., & Liang, Y. (2010). L1/2 regularization. *Science China Information Sciences*, *53*(6), 1159–1169.

Zhao, Q., Shen, L., & Ding, C. (2015). Dynamic spectrum access with goe-location database: A utility-based distributed learning approach. *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 918–922. https://doi.org/10.1109/ChinaSIP.2015.7230538

Zheng, Z., Shi, L., Sun, L., & Du, J. (2020). Short-term traffic flow prediction based on sparse regression and spatio-temporal data fusion. *IEEE Access*, *8*, 142111–142119. https://doi.org/10.1109/ACCESS.2020.3013010

Author Information

Vladimir II Christian Ocampo	**Lawrence Materum**
De La Salle University, Philippines	De La Salle University, Philippines
Contact e-mail: *vladimir_ii_ocampo@dlsu.edu.ph*	Tokyo City University, Japan

To cite this article:

Ocampo, V. C. & Materum, L. (2022). Wireless channel availability forecasting with a sparse geolocation spectrum database by penalty-regularization logistic models. *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 21*, 39-45.