Institution: University of Kent

Unit of Assessment: 10: Mathematical Sciences

Title of case study: Transferring Bayesian Nonparametric Methods and Variable Selection for Budget Forecasting to a Major Water and Sewerage Services Provider

Period when the underpinning research was undertaken: 2012-2018

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Maria Kalli	Senior Lecturer in Statistics	2017-present
Jim Griffin	Professor in Statistics	2007-2018

Period when the claimed impact occurred: 2018-2020

Is this case study continued from a case study submitted in 2014? No

1. Summary of the impact (indicative maximum 100 words)

Research at the School of Mathematics, Statistics and Actuarial Science (SMSAS) at the University of Kent is helping Berliner Wasserbetriebe (BWB) to improve its allocation of resources for their water network maintenance through the transfer of modern Bayesian nonparametric methods. The project, led by Maria Kalli from SMSAS, has fostered relations between a UK HE institution and international business, and has led to direct foreign investment into the UK HE sector. The knowledge transfer resulting from this project is evidenced by Kalli's improved budget-forecasting model. The accuracy of her model was proven during the test phase: the standard deviation (model v reality) has been reduced from up to 10% to less than 5% in the latest simulation (and to 0% in some cases). More accurate forecasting enables BWB to keep its tariffs stable. As BWB is the only company to provide water and sewerage services for Berlin and Brandenburg, stable tariffs will positively affect millions of people in the long term. This is the first project to make use of Bayesian statistics at BWB. It has raised BWB's awareness in this area and has led them to appoint a Data Scientist to work on this project and beyond.

2. Underpinning research (indicative maximum 500 words)

Berliner Wasserbetriebe (BWB) has a large network of pipes and invests almost €200 million annually to manage it. Investment forecasting is an essential task for managing budgets, tariff calculation, and the use of credit lines. Tariff calculation is highly regulated in Berlin and scrutinised by the state government, hence the necessity for high accuracy and low deviation in budget forecasting. Kalli has been working with BWB's planning department on a project that has developed a model that can accurately forecast the total cost of their water network (sewer and clean pipes) construction and reconditioning projects over the financial year. The work is based on the adaptation of the Bayesian nonparametric vector autoregressive (VAR) model **[R1, R2]**, which was developed by Kalli and Griffin from 2012 to 2018.

In **R1**, Kalli and Griffin developed a new Bayesian nonparametric model, called the BayesNP-VAR model, which accounts for non-linearity, non-constant mean and variance, and non-Gaussian innovations through directly modelling the joint and conditional distributions of the time series variables using infinite mixtures. In **R1**, the weights of the mixture depend on observed lagged values, allowing component transition densities to be favoured at different time periods. The BayesNP-VAR was shown to be a flexible model that is able to account for nonlinear relationships as well as heteroscedasticity in the data. Furthermore, the BayesNP-

VAR model was applied to US and UK macroeconomic time series, and superior predictive performance was observed when compared to existing Bayesian VAR models. **[R1]**.

R2 developed a prior for dynamic regression models that accounts for both time-varying regression coefficients and time-varying sparsity. In regression models with a large number of predictors, it is common to assume that only a subset of them is important for forecasting. In dynamic regression (DR) models, this assumption is naturally extended to subsets of important predictors that change over time. DR models have two forms of sparsity: 1) the proportion of predictor effects that are set to zero at time t; and 2) the proportion of time that subsets of predictors effects are set close to zero. The new Normal Gamma Autoregressive (NGAR) prior proposed in **R2** accounts for both of these forms of sparsity. The prior was applied to two forecasting problems, inflation forecasting and prediction of financial asset returns, and the NGAR led to better forecasts and at the same time allowed the identification of periods when predictor effects are very close to zero and so are effectively removed from the model.

The BWB collects both static and longitudinal data for each of their projects. Static data refers to factors that can affect the cost and timing of the project, such as: the district, the environmental zone, the level of priority, the method used to recondition the types of pipe and the lengths of the pipes used. Longitudinal data includes: estimated dates for the project to enter the tender process, the date of the successful tender, the estimated start dates, etc. The second layer of complexity in their data is the fact that each project cost is made up of several orders, and each of these orders has its own static and longitudinal factors. The attraction of the model developed in **R1** is its ability to identify changes in economic regimes. That is, when the economy will contract and expand, or be in a stable state. These regime changes are informed by values of covariates (factors) included in the model.

In a joint project with BWB, the BayesNP-VAR model was adapted to the requirements and the data of BWB. The new BWB model can predict when the project will move from one stage to the next, with each move informed by covariate values. Given the large number of covariates (factors at both the project and the order level), there is the danger of 'over-fitting'; that is, making the model overly complicated, which then leads to unreliable estimates and forecasts. To avoid this, Kalli used the ideas of sparsity developed in **R2**, which allows for the identification of the meaningful covariates that contribute to the estimation of the cost and the moves of the projects from stage to stage. The NGAR prior introduced in **R2** was adapted to accommodate the irregular time intervals and idiosyncrasies of each project.

The result is a BWB model that can provide project cost forecasts as well as project moves over time. This is crucial in the case of BWB, since this enables the company not only to build an accurate budget forecasting model, but also to predict stage progression for individual projects.

3. References to the research (indicative maximum of six references)

[R1] Kalli, Maria, and Griffin, Jim E. (2018). 'Bayesian nonparametric vector autoregressive models'. *Journal of Econometrics* 203(2): 267-282. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2017.11.009

[R2] Kalli, Maria, and Griffin, Jim E. (2014). 'Time-varying sparsity in dynamic regression models'. *Journal of Econometrics* 178(2): 779-793. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2013.10.012

4. Details of the impact (indicative maximum 750 words)

Berliner Wasserbetriebe (BWB) is the sole company providing water and sewerage services for the Berlin area. It has over 4 million consumers and is responsible for a network of pipes of around 19,000km. The company invests €192 million per year in its network.

At the 8th meeting of European Seminars in Bayesian Econometrics (October 2017), Kalli gave a talk on her ongoing research. This talk became the incentive for valuable discussions with BWB and led to direct foreign investment with a €100,000 contract being signed between BWB and SMSAS (Kent) in 2018 **[b]** for the new Bayesian nonparametric vector autoregressive models from **[R1]** to be extended and adapted to develop a robust forecasting model for determining the amount to be invested in plant projects and projects regarding the rehabilitation of the sewerage network.

The collaboration between BWB and SMSAS has resulted in better forecasting models for BWB's individual projects. BWB benefited from knowledge transfer, now being aware of the potential of Bayesian techniques and an ability to apply them for forecasting **[a]**. The company has also appointed a Data Scientist in its Planning and Construction department to further develop the use of Bayesian models in its network maintenance management **[a]**.

Kalli has developed a Matlab App (interface) of the forecasting model, which was being tested in late 2020, and BWB plans to 'use [it] for approximately 1,000 projects every year' **[a]**.

Kalli's model for BWB (and App) can be used to simulate where to allocate resources, which permits more accurate estimates. This is an appreciable improvement on the previous model that was used by BWB: the standard deviation (model v reality) is being reduced from up to 10% to less than 5% (the agreed threshold between BWB and SMSAS) in the latest simulation. For some specific networks, the latest simulations have out-performed the agreed threshold. Strohschein (Head of Asset Management at BWB) and Boldin (Project Manager) confirmed that 'the model developed by Kent researchers has a (model v reality) deviation of 0% for drinking water pipes' for their August 2019 budget forecast **[a]**. The initial testing phase performed by BWB also confirmed that 'results are promising and show increased accuracy in their budget forecasting' **[a]**.

In addition to its original forecasting function, the model and the App perform the following functions:

1. Make it possible to see in which stage every order is at the time the forecast is made, as well as when it is predicted to reach the next stages.
2. Include a feature that makes it possible to recreate old data to produce forecasts for previous years.
3. Implement a feature that allows new variables to be added, allowing the user to specify whether the variable is nominal, ordinal, or continuous.

These extra features have been made possible thanks to the techniques developed in **[R2]**. Sparsity allows for the new model to predict at which stage of the process each project is. This is a new feature for BWB, and Strohschein and Boldin confirmed that this new feature is 'promising' – 'the stage predictions from August up until September were correct for 80% of all the orders' **[a].** This has implications for the management of individual projects, as 'the construction start is a huge determining factor as well as a large factor of uncertainty' **[a]**. This will help the management department to check with engineers whether appropriate budgets are in place for each individual project on a monthly basis.

In Summary, Kalli and Griffin's research (**[R1]** and **[R2]**) has led to the design of a new and improved forecasting model for BWB through a €100k innovation contract (2018-21). This is a major project for BWB, as the model resulting from this project will be used for around 1,000 projects every year. This will ultimately benefit the 3.7 million inhabitants in the Berlin area, as the improved forecast of the network maintenance costs enables BWB to use its resources more effectively and keep tariffs stable (as opposed to increasing them). This collaboration already had wider implications; in particular, it has led to a change of practice at BWB. Strohschein and Boldin confirm that 'working with SMSAS has advanced their knowledge and thinking' **[a]**, and that 'they are already seeing benefits thanks to applying advanced Bayesian statistical modelling techniques in test phases' **[a]**. BWB is now fully aware of the potential of

Bayesian methods and have appointed a Data Scientist to use Bayesian models more widely in their management of the water and waste water projects.

5. Sources to corroborate the impact (indicative maximum of 10 references)

[a] Letter from Berliner Wasserbetriebe (BWB), providing full details on the impact of its collaboration with Kalli and Griffin.

[b] Contract (€100,000) and invoices for the completed project stages in the collaboration between the Kent team and Berliner Wasserbetriebe.