Research and Data Centre

Forecasting Algorithms

Technical annotation to Cahier 2024-4

Summary

Dutch institute for Justice and Security

Forecasting Algorithms

Technical annotation to Cahier 2024-4

Summary

F. ter Braak
M.S. Bargh
D.E.G. Moolenaar
B. Tims
R. Choenni

Cahier
This series comprises overviews of studies carried out by or for the WODC Research and Data Centre. Inclusion in the series does not mean that the sheet's contents reflect the viewpoint of the Minister of Justice and Security.

Summary

Policymakers would like to gain more insight into the (social) costs of crime, law enforcement and conflict resolution. Therefore, they would like to understand future trends in these fields so that the best possible decisions (in terms of policy and financial costs) can be made. Forecasting models may be used for this. To this end, the Forecasting Model for the Justice System (FMJS) was developed some time ago. This report examines the feasibility and usefulness of modernizing the FMJS using new developments in data processing and algorithms.

Forecasting Model for the Justice System

The forecasts about the numbers of recorded crime, suspects and everything that follows, and conflict resolution are currently made with the FMJS. This model covers virtually the entire Dutch judicial system, including (police) investigation, prosecution and sentencing, sanctions, subsidized legal aid in criminal cases and victim cares. In addition, the model also includes civil justice, administrative justice, legal aid in civil and administrative cases and the detention of illegal immigrants. The FMJS uses demographic, social and economic forecasts to forecast the inflow at the beginning of the system, such as recorded crime. These forecasts are then used to make forecasts of the inflow of cases at the Public Prosecution Service, and thus forecasts for the inflow at the courts and subsequently forecasts for the required sanction capacity. The FMJS is a combination of structural models, stock-flow models and time series models and includes approximately 6,600 equations. The parameters of the theoretical model are estimated using regression analysis on annual data. The result is an empirical model with which forecasts can be generated. The generated forecasts serve as a basis for a large part of the budget of the Ministry of Justice and Security.

Previous external evaluations show that the FMJS is well constructed and that users do not need a radically different model. However, the current FMJS was designed in a period when the availability of microdata was limited and a number of techniques were often theoretically known, but could not be implemented due to limitations in computer technology. It is therefore useful to investigate whether there have been developments in recent years in the field of data and forecasting techniques that can provide more or different insights.

Limitations

The purpose of the FMJS is to make estimates of the capacity needs of the Dutch judicial system. The FMJS assumes that the entire capacity requirement can be financed. There are therefore no budget restrictions in the FMJS. The FMJS also takes system effects into account. Only amounts are forecasted, such as number of cases or number of suspects or number of sanctions. The FMJS does not include prices. The FMJS forecasts the items on which judicial organizations are financed. This varies per organization and is determined by the organizations themselves in consultation with the Ministry of Justice and Security and not by the FMJS. The FMJS follows these decisions. If the method of financing is adjusted, the FMJS will be adjusted. A condition for the FMJS is that the criteria on which finances are provided, are quantifiable and

measurable. Because the method of financing is a decision made outside the FMJS, this report will not discuss it further. The estimates themselves and their accuracy are also not discussed here. These can be found in other Research and Data Centre (WODC) reports. The focus of this report is on the models that can be used to forecast trends for (long-term) strategic purposes and not on predictive models for operational or forensic purposes, such as 'predictive policing' or 'predictive sentencing'.

Research question and requirements

During the 2019 budget debate, the Minister for Legal Protection promised that he is willing to take another look at the FMJS. As a result, this report investigates to what extent new developments in the field of data analytics and machine learning could be used in the FMJS. Particular attention is paid to a number of aspects:

- Interpretability of the algorithm. How easy is it to explain in simple terms how the algorithm works? In other words, how intuitive is the algorithm?
- Complexity of the algorithm. How mathematically complex is the algorithm?
- Implementation. How much work does it take to implement the algorithm?
- Domain knowledge. Is it possible to use domain knowledge in combination with the algorithm?
- Network consistency in the justice system. Is it possible to use an algorithm to create a network-consistent model, that is, a model in which the outflow of one partner constitutes the inflow for the next partner?
- Time-component. Is it possible to include a time component in the algorithm? That is, can the algorithm dynamically make a forecast for the (medium and) long term?
- Noisy data. Can the algorithm deal with noisy data?
- Aggregated data. Is the algorithm applicable to aggregated data or to micro-data? And in the case of micro-data, can the results be aggregated?
- Computation time. How much computation time does it take to make forecasts?
- Interpretability of the model. Is the model easy to interpret and explain in simple terms?
- Fairness. To what extent can unfortunate choices or decision rules lead to the algorithm unintentionally becoming discriminatory?

This report mainly looks at forecasting methods not currently used for FMJS and to what extent they could improve the model without negating its current advantages, in particular network consistency. Therefore, any alternative algorithms must meet a number of requirements:

- The model must be network consistent. The forecast of the output of one partner must influence the forecast of the inflow of the subsequent partner.
- It must be possible to forecast seven years in advance, i.e. the budget horizon plus the years between the last known realization year and the first budget year.
- The model must be explainable in non-technical terms. Policymakers want to be able to understand why the forecasts are the way they are. In practice, this means that they must be traceable to specific input variables and that the estimated relationship must contain a certain degree of logic.
- Due to the planning of the budget process, the parameters of the model must be updated annually by mid-November. Because some data is only available at the end of September, this means in practice that the update must take place within a period of approximately six weeks.

- The chosen algorithm must be fair. Choices made or decision rules may not unintentionally lead to the algorithm becoming discriminatory in nature.

Methods

The methods we will discuss in this report can roughly be divided in four categories:

- methods that estimate the parameters of the current FMJS in a different way;
- methods that offer a different specification of (parts of) the model;
- methods that involve greater utilization of the data set with which the models are estimated and tested;
- combining methods or samples.

Examining the current FMJS

Linear regression is a simple and therefore frequently used class of algorithms, also in the current FMJS. Linear regression is also applicable if the data has been transformed in such a way that the transformation represents a linearization of a non-linear relationship. The current FMJS consists largely of linear regressions on transformed data, where the resulting equations are a linearization of a non-linear production function. If the data does not contain any significant problems, ordinary least squares is a good linear regression algorithm. If there are problems, a modified algorithm can be chosen depending on the type of problem. But there is currently no evidence to suggest that outliers, poorly measured or intercorrelated exogenous variables, truncated or censored data, and small or skewed samples form a major problem. Thus, there is no need for a modified algorithm that addresses these kinds of problems. A direction that is worth investigating further is elastic net regularization, and especially ridge regression. This method imposes a penalty on excessive complexity of the model.

Different specification of (parts of) the current FMJS

To forecast amounts of some sort, both linear and non-linear regression can be used, such as linear or non-linear time series or (non-linear) survival analysis Time series analysis is widely used for forecasting because it is relatively easy to implement and is integrated in many software packages. Linear time series analysis is already used to a limited extent in the current FMJS. However, the data must meet a number of conditions. For example, time series methods are mainly suitable for short-term forecasts, because time series models tend to revert to the mean of the process in the long term. This is a disadvantage when using time series for FMJS. Other forms of non-linear regression that may be of interest to the FMJS are algorithms that can handle duration data. Particularly, the (semi-) parametric survival methods seem promising for those parts of the justice field for which the duration is not known in advance, such as pre-trial detention or indefinite detention in a psychiatric prison hospital.

To forecast categorical outputs, classification algorithms are investigated, such as discriminant analysis, naive Bayes algorithm and logistic regression. Discriminant analysis is simple and computationally efficient. However, it is especially suitable for small samples with a limited number of independent variables. The naive Bayes algorithm is also not technically difficult, but the results are less intuitive if one has no knowledge of or affinity with probability distributions. This makes both algorithms less useful for the FMJS. For the FMJS, logistic regression seems to be a logical option to classify the many categories within the judicial systems such as: Should a suspect be held in remand custody? Should a suspect be prosecuted and/or trialled? What type of sanction should be imposed? These are typical choices that can be predicted with a logistic regression. Nevertheless, there are also limitations. The purpose of the FMJS is to forecast seven years ahead. It is difficult to include the time aspect in a logistic regression. Furthermore, logistic regression is applied to microdata which raises the question if the risk of privacy violations outweighs the increased performance of the model.

There are also algorithms that can forecast both numerical and categorical variables, such as k-nearest neighbours, decision trees, support vector machines and neural networks. These non-parametric algorithms have the advantage that few assumptions are made about the data in advance. Furthermore, the advantages and disadvantages differ per algorithm. All things considered, k-nearest neighbours and decision trees could be a good addition to the FMJS, although the main disadvantages of not being able to trace back to specific background factors or the absence of the time component, respectively, will have to be weighed against the advantages. Support vector machines and neural networks are less suitable mainly because the computation time of these algorithms is long and the updating of the models cannot be achieved in a responsible manner within the available time (approximately six weeks). Another disadvantage, particularly with neural networks, is that everything is interconnected and the model is difficult to interpret.

Greater utilization of the data set

There are several methods to improve the use of the data set. For example, instead of individual independent variables, combinations of independent variables can be included in the model, for example by means of principal component analysis or factor analysis. The major disadvantage is that the resulting forecasts cannot be traced back to specific background factors and therefore cannot be interpreted. Therefore, principal components analysis and factor analysis are less desirable for the FMJS.

Another way to improve the use of the data set is to reduce the validation errors, or the standard errors and the confidence intervals of the estimated parameters. By means of bootstrapping or cross-validation, multiple subsets can be drawn from the same sample, with which the same model is estimated each time. These methods are used to compare different models with each other in terms of predictive power. With a large number of models and/or observations, this can be very computationally intensive. Moreover, both bootstrapping and cross-validation are somewhat more difficult to apply to time series. Partly because the FMJS is usually not assessed solely on predictive power, but on several criteria, including theoretical correctness and interpretability. The marginal advantage of better predictive criteria and knowledge about the distribution of the parameters do not seem to outweigh the disadvantage, especially in cross-validation or the greater computational intensity. Bootstrapping can be interesting, especially in combination with other methods.

Combining methods or samples

There are a number of ways in which different forecasts of the same random variable can be combined. Bagging and boosting combine forecasts from the same algorithm but different samples. Boosting is less interesting because it requires very high data quality without outliers and noise and it is computationally intensive, meaning that the forecasts cannot be updated within the available time. But bagging, or bootstrap aggregating, could be interesting for the FMJS, because it is relatively easy to implement and the model itself does not become too complex, so that the results are still explainable. Bagging is also an essential part of the random forest algorithm. Random forest is a forest of decision trees, where each decision tree is built on a different bootstrapped sample of observations and/or a different random selection of the independent variables.

Ensemble averaging and stacking are methods that combine forecasts from different models trained with the same sample. With ensemble averaging, a weighted average is calculated, while with stacking a metamodel is developed. The disadvantage of the latter method is that it takes a lot of computational time and requires a lot of data. But ensemble averaging is a promising technique for the FMJS, because it is relatively easy to implement and different algorithms are often tried out in the test phase anyway. Until now, one model was ultimately chosen based on various criteria, although the differences in forecast quality were often minor. Alternatively, the average of the forecasts from multiple models could be calculated. For example, the inflow into the Public Prosecution Service could be forecasted based on the number of suspects registered by the police, but also by means of a time series analysis. The final forecast then becomes the (weighted) average of the outcomes of both algorithms.

Conclusion and recommendations

The most promising development direction for the FMJS seems to be causal/explainable artificial intelligence, but this is currently still an area of knowledge under development. Given the nature of the data, the purpose of FMJS and the preconditions, the most promising alternative algorithms are:

- elastic net regularization combined with linear regression;
- survival analysis;
- logistic regression;
- random forest;
- k-nearest neighbours;
- bagging;
- ensemble averaging;
- time series analysis for ensemble averaging.

These algorithms largely meet the requirements, but sometimes concessions will have to be made. In a follow-up study, a number of pilots will be carried out with these algorithms to see whether these algorithms actually lead to better forecasting performance.

The WODC (Research and Data Centre), a Dutch agency in the field of Justice and Security, is an independent knowledge institute that falls under the Dutch Ministry of Justice and Security. The WODC contributes to upholding and improving the rule of law by carrying out high-quality scientific research (or commissioning others to do so on its behalf), as well as by presenting solicited and unsolicited knowledge, points for improvement and (where possible) thinking strategies.

More information:

www.wodc.nl