Forecasting patent filings at the European Patent Office (EPO) using compositional data analysis techniques

Peter Hingley
European Patent Office
Munich

phingley@epo.org

A dynamic log-linear (DLL) regression model is fitted to annualised time series data on numbers of patent filings, in order to forecast future numbers of Total filings at the European Patent Office (Hingley and Park 2015, 2016). The model considers 28 source countries with independent variables for Gross Domestic Product, Research & Development Expenditures and autoregressive terms. A breakdown of Total filings according to industrial or technical areas of the underlying inventions induces additional compositions of the data to the breakdown by countries.

From the compositional data (CoDa) analysis perspective, Pawlowsky-Glahn (2013) and Coenders et al. (2015) have considered treatments for models of compositions with a total. These studies mainly involve geometric means from which a kind of total is obtained by multiplying by the sample size.

Here the additional forecasting power for modelling Total filings by the DLL model is assessed by considering a breakdown of the data according to three major industrial areas (IAs): *Electricals*, *Chemicals* and *Traditionals*. The IAs *Electricals* and *Traditionals* have grown markedly since year 2000, while *Chemicals* has grown more slowly. A model that includes the IAs is primarily useful for budgeting only insofar as it gives improved forecasts for the arithmetic total. Accumulation of results from separate DLL models fitted to each IA was not especially useful, although the individual forecasts for each IA are interesting enough.

As an alternative, CoDa related terms were added to a DLL model for Total filings (Hingley, 2016, pp. 27-31). Two isometric log-ratio (ilr) terms were included, that relate to the log proportions of the geometric means for the first two IAs. No ilr term for the third IA was included in order to avoid collinearity when fitting the regression model. This gave an improved least squares fit with a statistically significant ilr term for *Electricals*.

Several other forecasting models are routinely used that include simple models that sometimes produce better forecasts. A conventional CoDa based straight line regression model was also fitted according to the three IAs (ignoring country effects and independent variables except time). The total can be included again here via ilr terms and the forecasts contrasted with results from the DLL model. Some other issues involving modelling the total with compositional data will also be considered.

References

Coenders, G., Ferrer-Rosell, B., Mateu-Figueras, G. and Pawlosky-Glahn, V. (2015). *MANOVA of Compositional Data with a Total*. Article no 6 in http://compositionaldata.com/codawork2015/images/ProceedingsBook.pdf

Hingley, P. (2016). *Forecasting Total Numbers of Filings at the European Patent Office (EPO):- How useful are Breakdowns by Technologies?* International Institute of Forecasters Workshop on forecasting New Products and Technologies, Milan, 12 – 13 May 2016. http://www.innovazionesistematica.it/index.php?option=com_content&view=article&id=107&Itemid=1

Hingley, P. and Park, W. (2015). *A Dynamic Log-linear Regression Model to forecast Numbers of Future Filings at the European Patent Office*. World Patent Information, 42, pp. 19-27.

Hingley, P. and Park, W. (2016). *Do Business Cycles affect Patenting? Evidence from European Patent Office Filings*. Technological Forecasting and Social Change, article in press.

Pawlowsky-Glahn, V., Egozcuea, J. and Lovell, D. (2013). *The Product Space T (tools for compositional data with a total)*. Statistical Modelling 15(2), pp.175–190, & www.statistik.tuwien.ac.at/CoDaWork/CoDaWork2013Proceedings.pdf