International Journal of Forecasting 33 (2017) 502–512

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting compositional time series: A state space approach

Ralph D. Snyder[a], J. Keith Ord[b,*], Anne B. Koehler[c], Keith R. McLaren[a], Adrian N. Beaumont[d]

[a] Monash Business School, Monash University, Clayton, Victoria 3800, Australia
[b] McDonough School of Business, Georgetown University, Washington, DC 20057, USA
[c] Farmer School of Business, Miami University, Oxford, OH 45056, USA
[d] School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia

ARTICLE INFO

Keywords:
Log ratio transformation
Market shares
Maximum likelihood estimation
Model invariance
Multi-series models
New products
Prediction distributions
US automobiles sales
Vector exponential smoothing

ABSTRACT

A framework for the forecasting of composite time series, such as market shares, is proposed. Based on Gaussian multi-series innovations state space models, it relies on the log-ratio function to transform the observed shares (proportions) onto the real line. The models possess an unrestricted covariance matrix, but also have certain structural elements that are common to all series, which is proved to be both necessary and sufficient to ensure that the predictions of shares are invariant to the choice of base series. The framework includes a computationally efficient maximum likelihood approach to estimation, relying on exponential smoothing methods, which can be adapted to handle series that start late or finish early (new or withdrawn products). Simulated joint prediction distributions provide approximations to the required prediction distributions of individual shares and the associated quantities of interest. The approach is illustrated on US automobile market share data for the period 1961–2013.

1. Introduction

The need for forecasts of proportions arises in a wide variety of areas, such as business, economics, demography, and political science. Specific examples include the market shares of competing products, the proportions of jobs in different sectors of the economy, and the age composition of a population. In some cases, only measurements of the proportions are available; in others, such as market share data, both total sales and proportions are available, although the analysis typically focuses on the latter. Statistical methods for the analysis of data on proportions are known as compositional time series methods, and one essential component is some form of transformation to ensure that the specified random variables are non-negative and defined on a simplex so that they sum to one.

The monograph by Aitchison (1986) is a key reference for compositional data analysis. His analysis shows that we can draw on a wide variety of established and well-understood statistical methods by using the log-ratio transformation to map the proportions onto the real line. This transformation has become relatively standard over the years for both cross-sectional and time series analysis (Aitchison & Egozcue, 2005; Brundson & Smith, 1998; Quintana & West, 1988), and is adopted in our paper. A

* Corresponding author.
E-mail addresses: ralph.snyder@monash.edu (R.D. Snyder), Keith.Ord@georgetown.edu (J.K. Ord), abkoehler@gmail.com (A.B. Koehler), keith.mclaren@monash.edu (K.R. McLaren), abeaumont@unimelb.edu.au (A.N. Beaumont).

http://dx.doi.org/10.1016/j.ijforecast.2016.11.008
0169-2070/

key advantage of the transformation is that it enables us to develop the analysis on the whole real line, allowing parameter estimation to be based on least squares (minimum generalized variance) arguments. Of course, the construction of prediction intervals requires explicit distributional assumptions, with the multivariate normal distribution being the most common choice.

Once the log-ratio transformation has been applied, other forms of analysis become feasible, such as functional data analysis (cf. Hyndman & Booth, 2008), which enables the consideration of more general time-dependent mean and variance structures. This paper does not consider this possibility or the other approaches in the literature, based on the Dirichlet distribution (Grunwald, Raftery, & Guttorp, 1993) or the hyper-spherical transformation (Mills, 2010).

The novel feature of this paper is its coupling of the log-ratio transformation with linear innovations state space models and the associated technique of vector exponential smoothing (De Silva, Hyndman, & Snyder, 2009, 2010; Hyndman, Koehler, Ord, & Snyder, 2008). The rationale for this shift in emphasis lies in the nature of market share data, where new brands may be launched or old brands fade away. State space models can be structured easily to allow for the intrinsically non-stationary nature of a start-up, so that varying numbers of series ("births" and "deaths") may be considered. In contrast, vector ARIMA models (Barceló-Vidal, Aguilar, & Martín-Fernández, 2011; Brundson & Smith, 1998) assume that series are stationary after suitable differencing, and can notionally extend to the "infinite past". These authors do not consider births or deaths.

One issue with log-ratio time series models when there are three or more brands lies in the choice of the base series. When there are no restrictions on the parameter space, the results are not invariant to the choice of this base. One of the contributions of this paper is the identification of necessary and sufficient conditions for the forecasts to be invariant to the choice of base.

The emphasis in this paper is on pure time series models. Extensions that include explanatory variables are clearly possible, although as Brodie, Danaher, Kumar, and Leeflang (2001) observe, "causal models should only be used when input variables can be forecast with reasonable accuracy". In the absence of such information, they note that the naïve random walk model works well over short horizons, a scheme that one would hope could be bested by the systematic development of time series models. It should also be emphasized that we are not advocating the exclusive use of time series models for the analysis of market shares, but recognize that a combination of methods often adds value, as was demonstrated by Kumar, Nagpalb, and Venkatesan (2002).

The paper is structured as follows. The data on (grouped) company shares of the US automobile market, which serves as the example for our empirical work, are described in the next section. The basic model and the transformations that are used to ensure that the non-negativity and adding up constraints are satisfied are introduced in Section 3. Prediction distributions are developed in Section 4. Estimation procedures, with extensions to allow for series of unequal lengths due to births and deaths, are considered in Section 5, along with model selection issues and the construction of prediction intervals. A detailed analysis of market shares in the US automobile market is undertaken in Section 6, including two cases of new entrants into the overall market. Conclusions with discussion are provided in Section 7. The necessary and sufficient constraints on the parameter space that are required for base series invariance are derived in the Appendix.

2. Description of the data

The annual log-ratios of US total vehicle sales market share data (http://wardsauto.com/data-center), shown in Fig. 1, exhibit many of the issues considered in this paper and are used in Section 5 to illustrate our proposed approach. The original shares series related to 31 principal manufacturers who sell, or used to sell, in the American market. For the purpose of discussion, we consolidated them into six series as follows:

American: GM, Ford, Chrysler [3 series]

Japanese: Honda, Isuzu, Mazda, Mitsubishi, Nissan, Subaru, Suzuki and Toyota [1 series]

Korean: Hyundai and Kia [1 series]

Other (principally German): BMW, Daimler, Volkswagen and Other [1 series].

We selected these groupings with several factors in mind: they should be reasonably homogeneous in terms of market appeal; they should be small enough in number that they can be discerned in graphs; and they should not hide the entry of new products (Japan and Korea). It should be noted that any analysis that is based upon log-ratios will not be invariant under groupings of 'brands', but the effects will be minimized as long as we select reasonably homogeneous groups.

A fundamental issue concerning the use of log-ratio models needs to be addressed at this point. Hierarchical time series modeling (e.g., Hyndman, Ahmed, Athanasopoulos, & Shang, 2011) relies upon linear structures for combining elements (e.g., combine different options to define the sales of a particular model of automobile, then aggregate across all models of car produced by the manufacturer, then possibly across cars and trucks, and so on). In principle, the granularity of the data can be made as fine or as coarse as the analysis requires. However, the level of aggregation cannot be changed once the log-ratio transformation has been applied. Thus, the hierarchical and compositional approaches are complementary, in that hierarchical methods can be used to forecast component series within a given level of aggregation, and compositional methods are then used to compare entities at that level of aggregation.

The non-stationary nature of the transformed series is evident from Fig. 1. The market share for GM has declined steadily, meaning that the relative market shares of the other manufacturing groups have increased. The Japanese and Korean manufacturers have increased their shares relative to GM. The series for Chrysler chronicles the relative decline and recovery of that company in the seventies and eighties, whereas the "Other" series (mostly German manufacturers) reflects a decline followed by a more recent resurgence.