Paper 199

From Prediction to Validation: Who understands the Past, Controls the Future (Time Series).

EDWARD ARONOV

ABSTRACT

A new model for data prediction and validation has been developed with applications for market research, clinical trials and quality assurance. The process decomposed into typical structural components , such as a trend, seasonals, week day and pure random fluctuations. These parameters estimated by a multistep iterative procedure that minimized the error of residuals. Predictions, derived from the data for the corresponding periods of the time (like this visit and the previous one for the clinical trial) are used with optimal weighting procedure . Expected samples were calculated for the targeted time frame of prediction and validation. Results of the data processing have proven efficacy of this algorithm.

INTRODUCTION

The Wall Street Journal wrote recently that one of the most expensive myths in many corporations is that increasing amount of data and speed of delivery will increase income. Actually corporations are overwhelmed with the data. The processes of decision making and quality assurance usually are not appreciated enough.

Published systems of prediction do not solve some substantial problems: in that they do not consider that output could be independent from any explanatory variable; and that the trend is actually nonstationary with the possible turning points in the development of the process. Further, the data from this year and from the previous years are not used in an optimal way. Investigation of the behavior of the nonstationary process has helped us to find more effective description of the random changes over time.

This project aims to develop a new unified approach to the problem of data prediction and validation, using nonstationary time series, that may be used to improve the process of market research, new drug development, quality assurance, and other applications.

We describe here the approach to tackle these tasks by means of decomposition of seemingly chaotic changes into some typical components and estimation of the parameters of these components. We show how to find the optimal way to integrate statistics for the current period of time and the previous one.

BACKGROUND

Supply/ demand data, though random, sometimes show well developed shape of some typical behavior even with a visual inspection: trend with some turning points, seasonal changes and dependence on relation to the day-of-the week. These components could be very small in the period of estimation, but some of them become dominant in the future, in the period of prediction, and the error of prediction will be affected by them drastically.

A substantial part of a social and business activity is not related to the input variables, because this process is generated by hidden variables that are not observable.

Random fluctuations, if they are correlated over time, could be predicted by a Box- Jenkins methodology (especially for the short term prediction). These components are predicted with moving average and autoregression time series model, as each next sample is predicted by some weighted combination of the preceding samples /1/.

The process of estimation is more reliable, even from the computational point of view, if it is multistep (called here 'consequent refinement'). When parameters of any particular component are estimated, we will continue to work with residuals and use them as the input to a new procedure. Dimensions of the model are kept low for every step of refinement (to avoid the risk of an ill- conditioned system, which is typical for multidimensional models).

We have two sources of information to predict the future: from the previous period of time (for this year) and the corresponding period of time (the last year). Any of these have some advantages and disadvantages. This year will be related to the pattern of the weather conditions and business and social activity that are close to the next period of time, especially for a short term prediction. For a longer periods of time, the errors of these estimations become increasingly greater, because now this process is not affected too much by the near past. The predicted process could be very different from the real one.

Data from the previous year produce an adequate description for the corresponding period in general, but what they describe could have nothing to do with this year, especially for short term changes.

The original data represented with two sliding windows from the data base: two months before the base-date for this year and two months

after the base-date. The same windows are applied for the corresponding period of time for the last year.

NEW MODEL

The main functions of this model are:

1) To estimate parameters of all defined typical components (trend, hidden periodicals, day-of-the-week changes and pure random fluctuations) /5/

2) To use residuals from the previous step as an input to the next step of estimation;

3) To use different periods of time of estimation for a short and a long term prediction;

4) To apply the same procedure of estimation to data for current year and for the previous year separately;

5) To use a two-channel model to find the optimal weights for this year and the previous to find the optimal weights for all these structural components from this year and the previous year separately.

6) To test the model through comparison of predicted and actual data and to estimate the error of prediction.

These procedures are discussed below step by step.

We use the following next representation of the process in time $Y(t)=S(t)+R(t)+E(t)$; $S(t)=T(t)+P(t)+W(t)+H(t)$; $M(R(t))=0$; where S(t)-regular component, E(t)- error of estimation, M(R(t))- mathematical expectation, T(t)- trend, P(t)-seasonals, W(t) - week-day changes, H(t)-holidays and special events, R(t)- random fluctuations.

The trend and seasonal components describe the main part of equilibrium between supply and demand (including number of clients, rates, policy of denials, factors of competition), the weather condition, and the special social events. Because these factors are mixed together, and some of them could be available for registration in the future (but now they are usually not), we could not find optimum equilibrium analytically. The trend should describe the turning point, because the process of supply/ demand could change the direction of its development.

Statistical analysis showed that day-of-the-week component is an important explanatory variable. Finding regression for all seven days-of-the-week often leads to ill-conditioned fundamental matrix with big error of solution (disregarding whether it is noticed or not by a user).

We improved this step by transforming the original data into seven supplementary data sets, one for each day-of-the-week. We then apply regression to each subset of the data. As a result, seven-dimensional model transformed to seven one-dimensional models. Seven estimations are merged to produce the combined model, related to the original process. Reliability of the model improved because now this system is well- conditioned.

We describe seasonal components by the sum of the basic periodical functions. From the point of view of the business activity we could expect seasonal variables with some periods, but all of them have parameters with unknown amplitudes and phases. The equations are transcendental ones. We expand these functions to the sums of sines and cosines and apply regression procedure to the model that is linear. As a result, we find magnitudes and phases for all these components as a unique solution.

The pure random component is what is left after we extracted all deterministic components. It predicted by means of Box- Jenkins methodology /3,4/ (with SAS/ ETS®). We could also use a moving average/ autoregression procedure with two-step: short term and long term averaging. Auto-correlation function often is a second order function.

Parameters of all components are estimated separately for the two periods of estimation; for two months before the base-date of this year and for the corresponding period of the previous year. As a result, we can use now a two-channel model for prediction.

We find optimal weights for these two channels: $U(t)=aU1(t)+bU2(t)$; $V(t)=\min(L(t)-U(t))2$; $U1(t)=\{P1(t),T1(t),W1(t)\}$; $U2(t)=\{P2(t),T2(t),W2(t)\}$;
Where a, b- vectors of weight coefficients, L(t)-actual sales for the period of time of this year, selected as calibration period, V(t)-predicted process for the same period of time, U1, U2 - predictions for this year and for the previous year for every pair of particular components (trend, day-of-the-week, seasonal, random); Pk(t), Tk(t), Wk(t)- seasonals, trend, weekday changes in period of estimation with index k=1 related to this year and k= 2 -to the previous year.
This approach prevents the error of long term prediction (more than one week) from unstable growing and outstanding spikes, that is typical for published procedures.

RESULTS OF MODELING

Predicted data in comparison with the actual data, were selected to find the error of prediction, by means of graphs and goodness of appropriate statistical procedures and criteria.

Evaluation of the algorithm of prediction has been done to find the reasonable compromise between various contradictions. To trace the high- frequency random fluctuations, the system could use Box-Jenkins methodology.

This approach will need more time for adjustment and investigation of the problem of stationarity and stability of the auto-correlation function over period of time, because it is sensitive to abrupt changes and outliers.

The periodic spikes traced by the algorithm with a small average error, but sometimes there is a noticeable phase difference for some samples. This has happened because this process is not really pure periodical, but quasi- periodical, that is described by the model with a mean- square approximation over period of time. As a result, the balance of errors is small (a few percents) around the trend. but absolute difference in one particular day could be much larger. We consider the moving bias error the most useful criterion of the goodness of the model for many cases. Predicted samples should be used for a data validation, as they compared with actual ones.

CONCLUSION

The new SAS® program improves the process of the data prediction and validation of the random process for market conditions. This program estimates parameters of typical components that describe the process: seasonal changes, weekday, trend with turning points. These parameters are found even if they are small, and masked by random fluctuations. They used here for short term and long term prediction.

Predictions have been calculated from the data for this year and from the correspondent period of the previous year, and the final prediction has been done as optimal mixture of these two channels. The process of automatic calibration has been developed. The system was tested for some typical applications with different criteria of the estimation of the error of prediction. The model was evaluated for market research. This system shows improvements accuracy and reliability over currently used methods (2-10 times).

ACKNOWLEDGMENTS

The author would like to express his appreciation to Helen-Jean Talbott for the helpful comments.

REFERENCES

1. Box, G.P. and Jenkins, (1970), Time Series Analysis, Forecasting and Control, San Francisco, Cambridge, London, Amsterdam, Holden Day Inc.

2. Johnson, J., (1972), Econometrics Methods, Second Edition, New York, McGraw-Hill Book Company.

3. SAS Institute Inc. (1988), SAS /ETS® User's Guide. Version 6, First Edition, Cary, NC: SAS Institute Inc.

4.Jacob, P. (1992), "A Time Series Approach to Modeling Daily Peak Electricity Demands", Information Delivery to Utilities Sector. SAS Institute Inc.

5. Aronov, E. (1994), "Development of an Adaptive Model for forecasting of non-stationary processes", Proceedings of the 7th Annual Conference NESUG. SAS Institute.

6. Aronov, E. (1997), "A New Adaptive Model for prediction with the time series", Proceedings of the 10th Annual Conference NESUG. SAS Institute

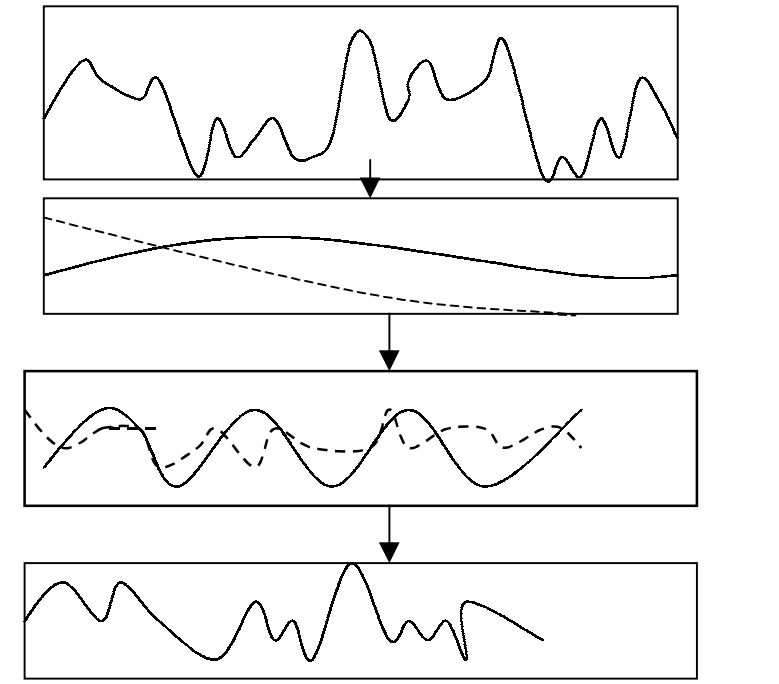
7. Lawson, C.L. and Hanson R.J., (1974), Solving Least Squares Problems, New Jersy, Prentice Hall, Inc. Englewood Cliffs.

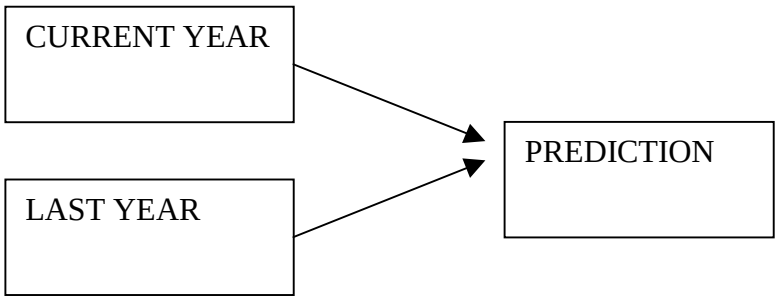
The author may be contacted at:
2528 Cruger Ave. apt. 4C
Bronx, NY 10467
Tel. 718-881-4153
E-mail: fractal271@aol.com

APPENDIX Illustrations

Fig.1. Structural decomposition of the original process of supply / demand. Short term trend, long term trend, seasonal changes, day-of-the-week, spikes, and chaotic changes represents original process. Parameters of this components estimated separately, step by step. The trend could be produced by summation of components with different parameters for different groups of clients and the different types of the business activity (segmentation).

Fig.2. Auto-calibration and the data window for the model of prediction. Parameters of components for this year and for the last year combined with optimal weight coefficients to produce the model for prediction. Selection of the data for the model of prediction. The period of estimation started 60 days before the base date for this year and corresponding period of time last year. Period of prediction ended 60 days after the base date for this year and the last year.




CURRENT YEAR
LAST YEAR
PREDICTION