What-If Analysis and Goal-Seek Analysis for Prescriptive Time Series Forecasting

Jane Chu[1], Jean Francois Puget[2]
[1]IBM Analytics, 71 S. Wacker Dr. 6th Fl., Chicago, IL 60606, USA
[2] IBM Analytics, 1681 Route Des Dolines, Les Taissounieres HB2, Valbonne, 06560, FR

Abstract
Many enterprises apply predictive analytics, e.g., transfer function models, Granger causality types of models, to time series data to forecast what might happen in the future. What can they do to turn forecasting into actionable insights? We propose two prescriptive techniques as answers: (1) what-if analysis predicts the possible outcomes based on different choices of actions; (2) goal seek analysis recommends the best course of actions with a desired outcome. Traditional prescriptive analytics can be hard to use as they (i) require users to specify their business problems as optimization models based on the predictive model, (ii) often require additional data. In this paper, we show how both analyses can be done by solving a constrained optimization problem in a system that combines predictive and prescriptive analytics. Moreover, both predictive and prescriptive models can be automatically derived from available history data plus user defined goals for outcome and constraints for actions. This results in a much more consumable form of prescriptive analytics.

Key words: goal seek analysis, Granger causality, predictive analytics, prescriptive analytics, transfer function model, what-if analysis.

1. Introduction

Many business and commerce metrics are recorded and stored as time series, which is a sequence of observations taken on equally spaced time points. Sales, expenses, earnings, customer satisfaction ratings, and other business metrics are observed over time, and the values of the metrics are recorded with an associated date or time.

The goal of time series analysis is usually to forecast future values, such as what future Sales values are for July to December, 2016. Because the auto-correlation nature of Sales series data, some statistical models, such as exponential smoothing, autoregressive integrated moving average (ARIMA), see IBM Inc. (2016b), etc. can be built based on its own past values to account for the historical patterns and used to forecast future Sales values. Such predictive models allow users to make good business decisions to account for the forecasts. For example, if Sales are forecast to drop in the next few months, then the business can budget appropriately, which is a proactive business decision to anticipate Sales decline. Because the model is based on the historical values that have been occurred, the forecasts of Sales cannot be changed.

On the other hand, some time series models, such as transfer function models, see IBM Inc. (2016b), Granger causality (which we call temporal causal modeling), see IBM Inc.

(2016a), etc., can be built to describe the causal relationships between the target series (Sales series) and predictor series (other series) to forecast future Sales values. Such predictive models can help users take better business actions to interact with the forecasts. For example, suppose that the number of Twitter mentions affects Sales. If Sales are forecast to drop in the next few months, then the business can change the future number of Twitter Mentions to cause Sales forecasts to increase. Because the model is based on other predictor series in addition to Sales' historical values, the forecasts of Sales might be changed due to some updated forecasts of Twitter Mentions.

There are two techniques when a user wants to interact with forecasts, i.e., change the forecasts of Sales. One is what-if analysis, and the other is goal seeking. Both options use the time series models with some predictors to generate new forecasts based on the user's desired outcomes.

In what-if analysis, a user manually sets the new forecasts for one or more predictor time series to see the repercussions in the target series. The analysis answers the question: How will my performance change if I control certain factors? For example, assume that the number of calls affect the monthly sales values. If this directed relationship is assumed true, user might change the forecasts for number of calls to see the effect on sales. What if I increase number of calls from 100 to 200 in September, 2016?

In goal seeking analysis, a user manually sets the new forecasts (goals) for the target time series to drive actions for predictor series. The analysis answers the question: What values of the predictor series will allow me achieve desired performance? For example, the user might set a goal for sales to see how forecasts for number of calls need to be changed to achieve it. What values of number of calls will allow me to reach sales of 700 (from 550) in September, 2016?

Goal seeking is prescriptive in that it provides recommended actions based on the user's desired outcomes, so it is natural to formulate it as a constrained optimization problem. While what-if is predictive in that it just predicts the new target values based on changes of predictors. However, when combining both analyses into a system, we will not force users to select one out of 2 analyses first. Instead, we allow users to freely specify any scenarios from what-if and/or goal seeking in the system. Each scenario would be formulated as a prescriptive model. Such a system is automatic in a sense that both predictive and prescriptive models can be automatically derived from available history data plus user defined goals for target/outcome and constraints for predictors/actions. This results in a much more consumable form of prescriptive analytics.

The rest of the sections are arranged as follows: Section 2 describes the automatic system that combines predictive and prescriptive analytics. Section 3 gives some scenarios based on a dataset while a few concluding remarks are in Section 4.

2. Automatic System

This paper proposes an automatic system that combines predictive and prescriptive analytics and generates solutions to satisfy user-specified goals and constraints. Users define goals and constraints for future dates and obtain metric values that are more likely to result in a scenario in which the targeted goals are met under constraints. The system includes the following steps:

1. Identify top predictors of a target metric based on the historical time series data.
2. Use the built models to forecast future values.
3. Allow users to conduct what-if and goal seek analyses by specifying goals and constraints in the forecasting period.
4. Solve the optimization problem that is generated from the goals and constraints.

We will describe these steps in the next four subsections in details.

2.1 Identify top predictors for the target metric

The first step is to identify top predictors and build a model for the target series, y_t, based on the historical data ($t = 1, \ldots, T$). Several different statistical techniques can identify predictors and build a model, such as temporal causal modeling, transfer function model (ARIMA + predictors), etc. The system can automatically check all possible predictors to identify the most important ones. Users could also manually select predictors, but it might be more efficient if the system does the selection automatically, in particular when the number of target series is huge.

Assume the system identifies the top K predictors, $X_{i,t}, i = 1, \cdots, K$, with the following temporal causal model:

$$y_t = \beta_0 + \sum_{\ell=1}^{L} \beta_{y,\ell} \cdot y_{t-\ell} + \sum_{i=1}^{K} \sum_{\ell=1}^{L} \beta_{i,\ell} \cdot X_{i,t-\ell} + \varepsilon_t \qquad (1)$$

where $\beta_0, \beta_{y,\ell}, \beta_{i,\ell},\ \ell = 1, \ldots, L, i = 1, \ldots, K,$ are parameters which need to be estimated; ε_t is an unobserved i.i.d. Gaussian error process with mean of zero and variance of σ^2 and L is the lag term for both target and predictors and it doesn't have to be the same for both target and predictors. However it is often set to be the same for the automatic model selection process when the number of target series is huge in practice.

2.2 Use the built models to forecast future values

The forecasts for y_t , $y_{t|T}$, at the current time T for several time points over the forecasting period, $T + 1, \ldots, T + h$, can be computed based on Equation (1) as follows:

$$y_{t|T} = \hat{\beta}_0 + \sum_{\ell=1}^{L} \hat{\beta}_{y,\ell} \cdot y_{t-\ell|T} + \sum_{i=1}^{K} \sum_{\ell=1}^{L} \hat{\beta}_{i,\ell} \cdot X_{i,t-\ell|T} \qquad (2)$$

where $y_{t-\ell|T} = y_{t-\ell},\ X_{i,t-\ell|T} = X_{i,t-\ell},\ t - \ell \leq T, i = 1, \ldots, K$. The forecasting values would be historical values if they happened before or at the current time T. On the other hand, the system will build models for predictor series to compute their forecasts, $X_{i,t-\ell|T}, t - \ell > T$.

The forecasting values based on historical data can be shown in Table 1:

Table 1: Historical and forecasting values for the target and predictors

Metrics \ Time	…	$T-1$	T	$T+1$	$T+2$	…	$T+h$
y_t	⋯	y_{T-1}	y_T	$y_{T+1\|T}$	$y_{T+2\|T}$	⋯	$y_{T+h\|T}$
$X_{1,t}$	⋯	$X_{1,T-1}$	$X_{1,T}$	$X_{1,T+1\|T}$	$X_{1,T+2\|T}$	⋯	$X_{1,T+h\|T}$

$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$X_{K,t}$	$\cdots$	$X_{2,T-1}$	$X_{2,T}$	$X_{K,T+1\mid T}$	$X_{K,T+2\mid T}$	$\cdots$	$X_{K,T+h\mid T}$

2.3 Allow users to specify goals and constraints

After building a time series model for the target series with some predictor series and forecasting future values, user can conduct what-if and goal seek analyses based on Table 1. Since our system can handle both analyses seamlessly, user is allowed to freely specify any goals for the target and constraints for the predictors in the forecasting period, but they cannot change any historical values for both the target and predictors because these have already happened.

Both goals and constraints can be entered one at a time based on the current time interval or be specified for the higher granularity than the current time interval. For example, assume the original measurements and forecasts are recorded and generated at a monthly level. The user can specify a goal value for the target or a constrained value for a predictor for a year or a quarter. Furthermore, both of them can be specified as a range. For example, the goal for the target, say Sales, for the next year increases at least 20% over the current year.

2.4 Solve the constraint optimization problem

The analysis that was defined by setting target goals and entering predictor constraints can be formulated as a constrained optimization problem. We need to find a solution that meets the goals of the target and satisfies the constraints of predictors by minimizing the change from the original forecasts.

First, let's define the decision variables to be solved in the optimization problem. Note that all variables might include both historical and forecasting periods.

$\tilde{y}_t$: Decision variable for target series at time t needs to be solved in the optimization problem, $t = T - L^* + 1, \cdots, T + h$;

$\tilde{X}_{i,t}$: Decision variable for the i^{th} predictor series at time t needs to be solved in the optimization problem, $t = T - L^{**} + 1, \cdots, T + h$; $i \in \{1, \cdots, K\}$ (because it is possible that some predictors aren't allowed to changed);

where $L^* = L^{**} = L$ for the temporal causal model in Equation (1), but L^* and L^{**} might be different if other time series model is used.

Then the constrained optimization problem can be formulated by specifying an objective function and several types of constraints. Under the optimization framework, the goals and constraints user entered are all considered as parts of following constraint types:

(1) Individual equality constraints: They are values user enters one at a time. To avoid the no solution issue with different types of constraints when solving the optimization problem, we will move them into the objective function.
(2) Grouped equality constraints: They are values user specifies for the higher granularity for either the target or any of predictors.
(3) Range constraints: They can be individual or grouped like equality constraints for either the target or predictors.

(I) The objective function:

Minimize the change between the original forecasting values and the changed forecasting values, which are the decision variables that need to be solved, of the target and predictors for all time points in the forecasting period:

$$\min_{\tilde{y}_t, \tilde{X}_{i,t}} \left(\sum_{t=T+1}^{T+h} w_{0,t} \left(\tilde{y}_t - g_{0,t} \right)^2 + \sum_{i=1}^{K} \sum_{t=T+1}^{T+h} w_{i,t} \left(\tilde{X}_{i,t} - g_{i,t} \right)^2 \right) \qquad (3)$$

where $g_{0,t}$ and $g_{i,t}$ are the objective value of $\tilde{y}_t$ and $\tilde{X}_{i,t}$, respectively. By default

$$g_{0,t} = \begin{cases} y_t, & t \le T \\ y_{t|T}, & t > T \end{cases} \text{ and } g_{i,t} = \begin{cases} X_{i,t}, & t \le T \\ X_{i,t|T}, & t > T \end{cases}, i = 1,2, \dots, K,$$

which will be overwritten if user defined a preferred value for $g_{0,t}$ or $g_{i,t}$, i.e., the individual equality constraints mentioned above. And $w_{0,t}$ and $w_{i,t}$ are the optimization weights. The purposes of setting these weights are two folds: (1) to make all the series on the same sale, (2) to assign a higher optimization weights onto those time points with user-specified objective values. A recommended set of optimization weights are:

$$w_{0,t} = \frac{\sigma^2_{y_{T+h|T}}}{\sigma^2_{y_{t|T}}} \cdot \frac{1}{\mu_0^2 + \sigma_0^2} \cdot a^{I\{g_{0,t} \text{ is user defined }\}}, \text{ and}$$
$$w_{i,t} = \frac{1}{\mu_i^2 + \sigma_i^2} \cdot b^{I\{g_{i,t} \text{ is user defined }\}}, i = 1,2, \dots, K,$$

where $\sigma^2_{y_{t|T}}$ is the forecasting variance of the target at time $t, t = T + 1, \cdots, T + h$. μ_0 and μ_i are the means of the target and ith predictor series, respectively. σ_0 and σ_i are the standard deviations of the target and the ith predictor series, respectively. The constants a and b default to 10,000 and 1,000, respectively.

(II) The constraints:

Equation (3) is solved subject to the following constraints:

(a) Grouped equality constraints:

$$\boldsymbol{c}_0' \tilde{\boldsymbol{y}} = G_0 \text{ and } \boldsymbol{c}_i' \tilde{\boldsymbol{X}}_i = G_i$$

where $\tilde{\boldsymbol{y}} = (\tilde{y}_{T+1}, \tilde{y}_{T+2}, \dots, \tilde{y}_{T+h})'$, $\tilde{\boldsymbol{X}}_i = \left(\tilde{X}_{i,T+1}, \tilde{X}_{i,T+2}, \dots, \tilde{X}_{i,T+h} \right)'$, $i \in \{1, \cdots, K\}$, and $\boldsymbol{c}_i$, $i \in \{0,1,2, \dots, K\}$ is a vector of length h made of 0s and at least two 1s. If there is only one 1 in $\boldsymbol{c}_i$, constraint (a) will reduce to individual equality constraint which is handled by the objective function instead of the constraints. This constraint is a user input. We provide two scenarios of constraint (a) to users:

(a.1) Grouped summation equality constraint

For example, suppose now is 2015-12 and we have monthly sales data as the target series with forecast length h = 12. User can set the summation of the sales in the first quarter of 2016 to be 5,000 as the goal. In this case, for constraint (a), $\boldsymbol{c}_0 = (1,1,1,0,0,0,0,0,0,0,0,0)'$ and $G_0 = 5{,}000$.

(a.2) Grouped average equality constraint

Use the same example in (a.1). User can set the average of the sales in the second quarter of 2016 to be 1,000 as the goal. In this case, for constraint (a), $\boldsymbol{c}_0 = (0,0,0,1,1,1,0,0,0,0,0,0)'$ and $G_0 = 1{,}000 \times 3 = 3{,}000$.

(b) Range constraints:

$$L_0 \le \boldsymbol{c}_0'\tilde{\boldsymbol{y}} \le U_0 \text{ and } L_i \le \boldsymbol{c}_i'\tilde{\boldsymbol{X}}_i \le U_i, i \in \{1, \cdots, K\}$$

where $L_i \in \mathbb{R} \cup \{-\infty\}$ and $U_i \in \mathbb{R} \cup \{+\infty\}$. Multiple range constraints can be applied to each series as long as they are not conflict with each other. This constraint is also a user input and we provide three scenarios for users to set the range constraints:

(b.1) Individual range constraints

With this scenario, user sets a range constraint for only one observation at a time. Use the same example in (a.1) with the product price as one of predictor series. User can set the goal of sales of 2016-02 to be 10,000 while set the lower limit of the price of 2016-02 to be $100. In this case, user is adding an individual range constraint to the price series at 2016-02, for constraint (b), $\boldsymbol{c}_1 = (0,1,0,0,0,0,0,0,0,0,0,0)'$, $L_1 = 100$ and $U_1 = +\infty$.

And user can add multiple individual range constraints to the same time series. In the same example, if user also does not want the price of 2016-01 goes below $90, s/he can add another constraint of type (b) with $\boldsymbol{c}_1 = (1,0,0,0,0,0,0,0,0,0,0,0)'$, $L_1 = 90$ and $U_1 = +\infty$.

(b.2) Grouped summation range constraints

With this scenario, user sets range constraints on the summation of multiple time points in the forecasting period. Use the same example in (a.1). If user sets the constraint that the forecasted total sales of 2016 to be greater than 10,000 and less than 20,000. Then for constraint (b), $\boldsymbol{c}_0 = (1,1,1,1,1,1,1,1,1,1,1,1)'$, $L_0 = 10{,}000$ and $U_0 = 20{,}000$.

(b.3) Grouped average range constraints

With this scenario, user can set constraints on the average of multiple time points in the forecasting period. Use the same example in (a.1). If user set the constraint that the forecasted average sales of the first half year of 2016 to be greater than 1,000 and less than 2,000. Then for constraint (b), $\boldsymbol{c}_0 = (1,1,1,1,1,1,0,0,0,0,0,0)'$, $L_0 = 1{,}000 \times 6 = 6{,}000$ and $U_0 = 2{,}000 \times 6 = 12{,}000$.

The constraints described in (a) and (b) are suitable for metric series with aggregation functions of sum or average (mean). If user doesn't specify clearly the constraint is summation or average, then it is handled based on the metric series' aggregation function.

(c) Historical data cannot be changed:

$$\tilde{y}_t = y_t \text{ and } \tilde{X}_{i,t} = X_{i,t}, t \le T$$

(d) All changed values are satisfied with the time series model (Equation (2)) used to compute the target forecasting values:

$$\tilde{y}_t = \hat{\beta}_0 + \sum_{\ell=1}^{L} \hat{\beta}_{y,\ell} \cdot \tilde{y}_{t-\ell} + \sum_{i=1}^{K} \sum_{\ell=1}^{L} \hat{\beta}_{i,\ell} \cdot \tilde{X}_{i,t-\ell}\,, \; t = T+1, \cdots, T+h$$

where $\tilde{y}_{t-\ell} = y_{t-\ell}$, $\tilde{X}_{i,t-\ell} = X_{i,t-\ell}$, $t - \ell \leq T, i = 1, \ldots, K$.

3. An Example

The data set used to demonstrate the system is monthly sales revenue (Revenue) of a product in a company from January, 2002 to September, 2011, along with other possible predictor series, such as foot traffic in the stores (Foot_Traffic), online ads (Online_Ads), TV ads (TV_Ads), direct mail offers (Dmail_Offer) and email offers (Email_Offer). The system would automatically select two top predictor series: Foot_Traffic $(X_{1,t})$ and Online_Ads $(X_{2,t})$ and the built model for Revenue (y_t) is:

$$y_t = y_{t-12} + 318.68\left(X_{1,t} - X_{1,t-12}\right) + 1.546\left(X_{2,t} - X_{2,t-12}\right).$$

The forecasts for the target and two predictor series for the next fiscal year (2011-10 to 2012-09) are listed in Table 2 while Figure 1 displays both historical and forecasting values of all three series.

Table 2: Forecasts for Revenue, Foot_Traffic and Online_Ads for 2011-10 to 2012-09

Time Metric	2011 -10	2011 -11	2011 -12	2012 -01	2012 -02	2012 -03	2012 -04	2012 -05	2012 -06	2012 -07	2012 -08	2012 -09
Revenue	23396	21781	29772	21766	18249	17643	17207	18629	17590	18219	30261	23072
Foot_Traffic	38.42	37.57	47.42	49.46	43.27	45.53	40.74	43.90	45.37	44.84	43.98	46.59
Online_Ads	11690	11393	15263	11821	12024	12176	12017	12234	12246	12096	12246	11576

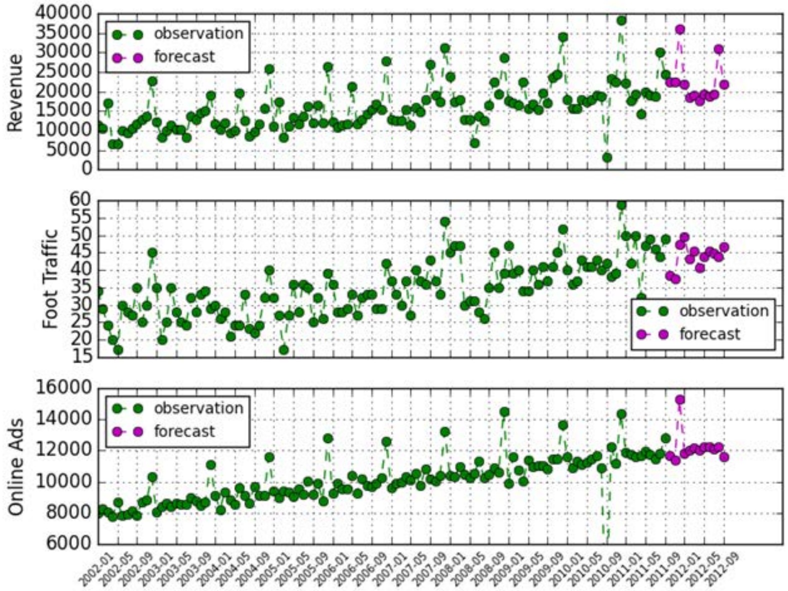

Figure 1: Observed and forecasting values for Revenue, Foot_Traffic and Online_Ads

Two scenarios are demonstrated. For Scenario 1, the goal is to increase the total forecasting revenue of next fiscal year (2011-10 to 2012-09) by 10% of the original forecasts. The solutions (optimized forecasts) and original forecasts for Revenue, Foot_Traffic and Online_Ads are shown in Figure 2.

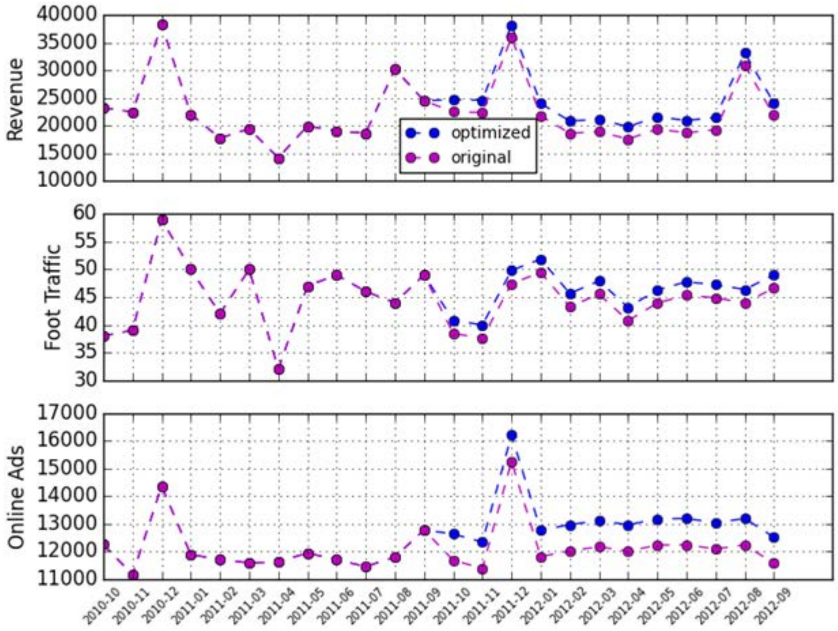

Figure 2: Optimized & original forecasts for Revenue, Foot_Traffic and Online_Ads in Scenario 1

For Scenario 2, the goal is the same as the above but Online_Ads from 2012-01 to 2012-09 cannot exceed 12,000. The solutions and original forecasts are shown in Figure 3.

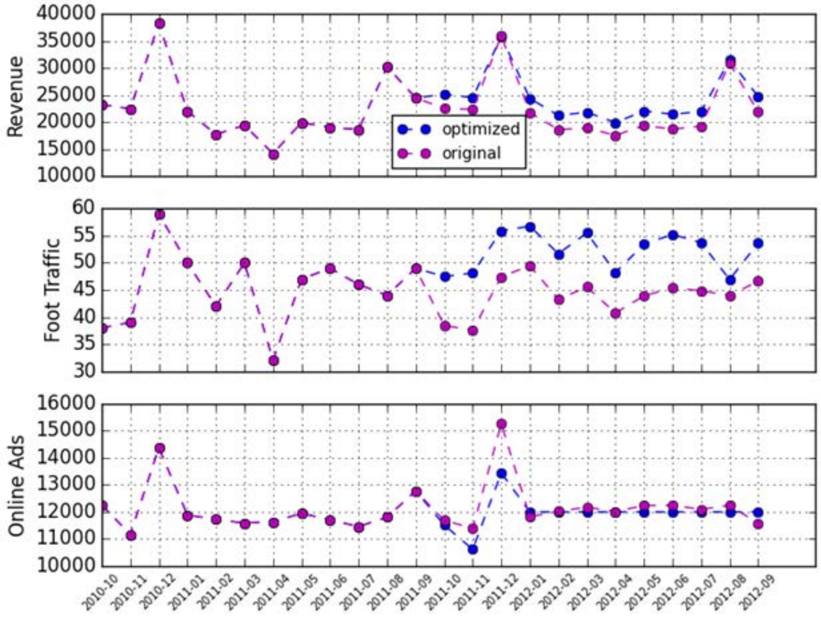

Figure 3: Optimized & original forecasts for Revenue, Foot_Traffic and Online_Ads in Scenario 2

4. Conclusion

The proposed system combines predictive and prescriptive analytics for time series data. It has several advantages over more traditional ones for cross sectional data. (1) The required input from user is minimal. (2) The objective function has the default form while allowing user to change to their own according to their business requirements. (3) It operates on time series data, so the solution is dynamic over the forecasting period. (4) It provides truly actionable results if predictor series can allow user to act on.

References

IBM Inc. (2016a), Temporal Causal Modeling Algorithms, *IBM SPSS Statistics 24 Algorithms*, Chicago, IL, 1064–1081.

IBM Inc. (2016b), TSMODEL Algorithms, *IBM SPSS Statistics 24 Algorithms*, Chicago, IL, 1118–1135.