Employee turnover forecasting for human resource management based on time series analysis

Xiaojuan Zhua, William Seaverb, Rapinder Sawhneya, Shuguang Jia[[1]](#footnote-1)\*, Bruce Holtc, Girish Upretia and Gurudatt Bhaskar Sanila

aDepartment of Industrial & Systems Engineering Department, University of Tennessee, Knoxville, USA; bDepartment of Statistics, Operations, & Management Science, University of Tennessee, Knoxville, USA; cIntegrated Planning, Y-12 National Security Complex, Oak Ridge, USA

**Acknowledgements**

We thank Minoo Mardookhy and Xiaocun Sun for valuable comments.

**Disclaimer**

This work of authorship and those incorporated herein were prepared by Contractor as accounts of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Contractor, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, use made, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favouring by the United States Government or any agency or Contractor thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency or Contractor thereof.

**Copyright Notice**

This document has been authored by a contractor/subcontractor of the U.S. Government under contract DE-AC05-00OR-22800. Accordingly, the U.S. Government retains a paid-up, nonexclusive, irrevocable, worldwide license to publish or reproduce the published form of this contribution, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, or allow others to do so, for U. S. Government purposes.

Employee turnover forecasting for human resource management

based on time series analysis

Xiaojuan Zhua, William Seaverb, Rapinder Sawhneya, Shuguang Jia[[2]](#footnote-2)\*, Bruce Holtc, Girish Upretia and Gurudatt Bhaskar Sanila

*aDepartment of Industrial & Systems Engineering Department, University of Tennessee, Knoxville, USA; bDepartment of Statistics, Operations, & Management Science, University of Tennessee, Knoxville, USA; cIntegrated Planning, Y-12 National Security Complex, Oak Ridge, USA*

In some organizations, the lead time of hiring is often long owing to responding to human resource requirement associated with technical and security constrains. Thus, the human resource departments in these organizations are pretty interested in forecasting employee turnover because a good prediction of employee turnover could help the organizations to minimize the cost and impacts from the turnover on the operational capabilities and the budget. This study aims to enhance the ability to forecast employee turnover. The research utilizes various statistical modelling techniques in time series to identify optimal models for effective employee turnover prediction. More than 11-year monthly turnover data is used to build and validate the proposed models. Compared with other models, the dynamic regression model with additive trend, seasonality, interventions, and external predictor is able to predict the turnover efficiently with training R2=0.77 and holdout R2=0.59. The forecasting performance of optimal models in this study confirms that time series modelling approach has a reasonable ability to predict employee turnover for various scenarios.

Keywords: human resource management; turnover; time series; forecast

# Introduction

Prediction of employee turnover is a topic that has drawn the attention of researchers and human resource managers because employee turnover cost impacts both the operational capabilities and the budget of an organization. Employee turnover is both costly and disruptive to the functioning of most organizations, and both private firms and governments spend billions of dollars every year managing the issue according to Leonard ([2001](#_ENREF_25)).Furthermore, at crucial times, organizations find themselves short of niche skill-sets and resources which require time and planning to acquire. The lead time for hiring is often long, particularly when special skills are involved, and in some organizations like U.S. national laboratories, due to the security clearance requirements and training, the process could take months. Therefore, a good prediction of employee turnover at firm and departmental levels is essential for effective planning, budgeting, and recruiting in the human resource field.

Human resource planning (HRP) is an ongoing process of systematic planning to achieve optimum use of the human resource pool in an organization. For an organization to execute their tasks efficiently and effectively, they need to ensure that the right people are available at the right places and at the right times to execute the tasks with the highest quality ([Khoong, 1996](#_ENREF_24)). Over the years, organizations have been able to scale up their efforts and success in manufacturing, marketing and financial plans. However, organizations have always struggled to develop sustainable HRP models ([Heneman, Schwab, Fossum and Dyer, 1993](#_ENREF_20)). The objective of sustainable HRP models is to ensure the best match between employees and jobs to avoid manpower shortages or surpluses ([Čambál, Holková and Lenhardtová, 2011](#_ENREF_10)). To achieve this balance employee turnover is an important metric that is often central to organizations workforce planning and strategy.

As summarized in Table 1, some of the previous studies attempt to identify the explanatory predictors of employee turnover. For instance, [Bluedorn (1982)](#_ENREF_6) found that the turnover model appears to be related to the individual’s perception of environmental opportunities, routine, age, and the length of service. [Balfour and Neff (1993)](#_ENREF_4) noticed that caseworkers with more education, less experience, and less stake in an organization are more likely to turnover. According to the research conducted by [Wright and Cropanzano (1998)](#_ENREF_41), they found that emotional exhaustion was associated with both job performance and subsequent turnover, but not related to job satisfaction. [Morrow, McElroy, Laczniak and Fenton (1999)](#_ENREF_26) used employee absenteeism and performance to predict employee turnover. The result from this study shows a positive correlation between absenteeism and voluntary turnover, and a negative correlation between performance ratings and voluntary turnover. The study conducted by [Thaden, Jacobs-Priebe and Evans (2010)](#_ENREF_35) indicated that organizational culture may potentially be an important factor to retain the workers in an organization.

Meanwhile, some other studies try to build turnover prediction models through techniques, such as regression, neural network (NN), and data mining. For example, [Ng, Cram and Jenkins (1991)](#_ENREF_28) used a proportional hazards regression (PHR) to develop turnover prediction model and their forecasts are much more accurate. In the study conducted by [Sexton, McMurtrey, Michalopoulos and Smith (2005)](#_ENREF_32), NN combining with modified genetic algorithm was used to build the prediction model for turnover. Alao and Adeyemo ([2013](#_ENREF_2)) applied decision tree skill on the employees’ demographical information and personnel records to identify attributes contributing to employee turnover.

In these studies, the data source was acquired either from human resource employment records and demographical information or survey with time horizon ranging from 1 to 28 years. Most of data is monthly data. It is worth mentioning that only few of these studies considered macro-economic factor in their prediction models, which could have significant impact on employee turnover.

Although there have been some efforts to predict employee turnover behaviour, no study has been conducted on turnover with time series forecasting skills, especially univariate time series forecasting, to predict employee turnover. In this study, the authors attempt to fill this gap. The advantage of time series forecasting approach is that it is not necessary to identify the determinants of turnover and helpful to evaluate the effects of either a planned or unplanned intervention ([Velicer and Fava, 2003](#_ENREF_36)).

Thus, this article has four parts. First, there is the introduction which covers the objective of the paper and a literature review. Second, there is a synopsis of tools for finding time series patterns and preparing the data for analysis as well as specific forecasting methods. The methods section is followed by the results and discussion of the study. Finally, there is conclusion of the results, practical implications, and limitations of the paper.

Table 1. Summary of previous research on employment turnover forecast.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Authors (Year) | Data Acquisition | Data Horizon | Methods | Software | Economic Indicator | Response Variable | Estimate | Model Evaluation |
| Bluedorn ([1982](#_ENREF_6)) | Employee records and Survey | 1 year | Correlations, multiple regression | N/A | No | Number | Point with intervals | R2=0.22,  Adjusted R2=0.11 |
| Ng et al. ([1991](#_ENREF_28)) | Survey | N/A | Hazard proportional model | BMDP 2L | No | Probability | Point with intervals | Pair t-test |
| Balfour et al. ([1993](#_ENREF_4)) | Employee records | 33 months | Non-linear logistic regression | N/A | No | Probability | Point | Chi-square values |
| Feeley et al. ([1997](#_ENREF_15)) | Survey | 60 months | Social network, logistic regression, correlation | NEGOPY, UCINET | No | Probability | Point | R2=0.23 |
| Wright et al. ([1998](#_ENREF_41)) | Survey | 1 year | Hypothesis test, correlation, logistic regression | N/A | No | N/A | N/A | Correlation r=0.34, P<0.01 |
| Morrow et al. ([1999](#_ENREF_26)) | Demographic information and employee records | 2 years | Logistic regression, correlation | N/A | No | Probability | Point | (-2 log likelihood) chi-square=193.13 |
| [Sexton et al. (2005)](#_ENREF_32) | Demographic information and employee records | 10 years (yearly) | NN | FORTRAN | Yes | Leave or not | Point | Type I error=0.25% Type II error= 5.83% |
| Hong et al. ([2007](#_ENREF_21)) | Survey | N/A | Logit and probit model | SPSS | No | Probability | N/A | R2=0.5, Quadratic Probability Scores = 0.18 for training and 0.12 for test |
| Nagadevara et al. ([2008](#_ENREF_27)) | Demographic information and employee records | 3 years | NN, logistic regression, classification/regression trees, discriminant analysis | N/A | No | Leave or not | Point | Contingency table |
| Thaden et al. ([2010](#_ENREF_35)) | Survey | 2 years | Multiple regression | N/A | No | Duration | Point with intervals | R2=0.56, P< 0.001 |
| [Größler and Zock (2010)](#_ENREF_16) | Employee records | 360 months | System dynamics | N/A | No | Number | Point | N/A |
| Saradhi et al. ([2011](#_ENREF_31)) | Survey | 2 years | SVMs, random forest, Naïve Bayes classifiers | N/A | No | Probability | Point | True/false positive rate and precision |
| Alao et al. ([2013](#_ENREF_2)) | Employee records | 28 years (yearly) | Decision tree | WEKA See5 | No | Probability | Point | True/false positive rate and precision |

# 

# Methods

# *Data sources and preparation*

The human resource data is provided by a large multipurpose research organization in the U.S. The dataset consists of over 8,000 observations. These observations are active and terminated employees’ demographic information from November 2000 to January 2012 including metrics such as payroll information, last hired date, termination date, age, years of service, gender, job classification, and department code. The turnover dataset is summarized by month and consists of 135 months spanning from November 2000 to January 2012.

U.S. Monthly composite leading indicator (CLI) data published by Organization for Economic Co-operation and Development (OECD) from November 2000 to January 2012 is employed as a predictor of cyclical component of employee turnover series, since, in practice, CLI is used to give an early indication of turning points in the macro-economic cycle. CLI data used for this study was constructed by OECD through aggregating 7 components together. These 7 components include the number of dwellings started, net new orders for durable goods, share prices-NYSE composite, consumer sentiment indicator, the average weekly hours worked by manufacturing workers, the purchasing manager index, and spread of interest rates ([Organisation for Economic Co-operation and Development, 2013](#_ENREF_29)). The authors have no information on weights of the index construction.

# *Pattern analysis, cross-correlations, and outlier identification*

To model a time series, it is important to look for patterns in the turnover series. First, this pattern analysis is simply addressed with a time plot of the series (a scatter plot over time) and box plots for the seasons or the months. In this case, the seasonal pattern of the turnover series is tested through Kruskal-Wallis and ANOVA tests (P<0.05), which do not correct for any trend in the series. The second stage of the pattern analysis is to use autocorrelation (ACF) and partial autocorrelation (PACF) plots to identify seasonal, autoregressive, and moving average patterns. If there are external variables (as in this case), the third stage of pattern analysis would be to examine the cross-correlations between (and external variables (CLI () over time). The cross-correlation function (CCF) is used to identify lags of CLI () that might be useful predictors of turnover series ( The longer the lag is the greater the forecast horizon of the model when using external variables.

All of the previous stages of the pattern analysis could be contaminated by outliers so it is important to identify outliers before fitting the actual forecasting model. Box plot analysis by season can be used to informally flag outliers (this approach tends to over-identify outliers) while ARIMA methods in conjunction with statistical process control (SPC) tends to correctly identify the number of outliers in a series. Conservative ARIMA models ((1,0,1) or (0,1,1)) are used in this control charting. The residuals from ARIMA(1,0,1) are divided by the square root of mean square error to standardized them and the outliers are identified with the value greater than ±3 standard deviations from zero by scatterplots of the standardized residuals ([Alwan and Roberts, 1988](#_ENREF_3), [Grznar, Booth and Sebastian, 1997](#_ENREF_17)). Identifying the outliers using SPC and then smoothing them is a way to refine the data for further analysis and to facilitate finding the underlying pattern in the data series. Smoothing outliers in a time series is to filter out random noise and other irregularities. The smoothing technique is based on the assumption that the data point at time *t* should notdeviate dramatically from the point at times *t*–*2s*, *t*–*s*, *t, t+s,* and *t+2s* with a smoothing window of 5 seasonal periods. When the outlier is identified, it is adjusted to be more similar to its neighboring points ([Grznar et al., 1997](#_ENREF_17)). In this study, unusual observations were smoothed using a nonlinear data smoothing method, based on repeated medians (RMD) of a five period span (as shown in equation (1) ([Velleman, 1980](#_ENREF_37)):

(1)

Where, is the actual smoothed value at time *t*, is value of response variable at time *t*, and *s* is the number of total seasonal periods which is 12. This smoothing is only done on potential outliers within their season not across adjacent periods. Sometimes outliers can distort normality, white noise, cross correlations, ACF and PACF, and the predictive performance of the model. Thus, dealing with outliers properly means going back and asking hard questions about these unusual observations from a human resource perspective, and that can truly add understanding and improved forecasting ability. It is always possible that the outliers are not truly unusual events but an intervention or a change point where early retirement incentives were offered, another company was purchased, or a section of the original company was sold. There are many possibilities; but if there is not an over-identification of outliers, a good human resource department should be able to provide quick answers on the unusual situations.

# *Time series analysis*

In time series forecasting, past observations of the same variable are collected and analysed to develop a model describing the underlying relationship. This modelling approach is particularly useful when little knowledge is available on the underlying data generating process or when there is no satisfactory explanatory model that relates the prediction variable to other explanatory variables ([Zhang, 2003](#_ENREF_42)). In this analysis, basically univariate time series methods are used to identify an optimum forecast with and without external variables. The statistical software used was Number Cruncher Statistical System (NCSS) and SAS. The data is analysed for two partitions of the data: training sample (November 2000 – January 2011) and holdout sample (February 2011- January 2012). For each model, the training sample is used to build the model and the holdout sample is used for the validation of the model because the most recent time series data is considered the most important factor for prediction purposes ([Bergmeir and Benítez, 2012](#_ENREF_5)).

### Univariate methods (without external variables)

These univariate time series analysis models use months and trend (for the most part) as predictors of turnover. The univariate models without external variables which were used in this study include time series regression, decomposition, Winter’s Exponential Smoothing (WES), and Box-Jenkins Autoregressive Integrated Moving Averages (ARIMA). One critical reason for using these kinds of time series models was so that there was not a constraint on the forecast horizon, i.e., how far in the future one could predict.

### Time series regression

The univariate time series regression model adopts trend and seasonality as two predictors. The additive time series regression model with intercept, trend, monthly seasonality, and error terms takes the form as shown in equation (2):

(2)

Where, is the value of response variable at time *t*, is the coefficients estimated by regression, is a continuous variable representing trend with value from 1 to *n*, is dummy variable representing seasonal periods. In addition, the additive regression models with interventions (pulse and step) are also considered in the analysis due to the downsize policy in certain year in our case. The multiplicative time series regression model with intercept, trend, monthly seasonality, and error terms is also considered with the form as shown in equation (3):

(3)

Where, is natural log transformation of. For these regression models, the significance of the model and the variables is examined by p-values at 0.05 significant level, lack of collinearity, good holdout performance, and valid regression assumptions.

### Decomposition

Decomposition time series methods attempt to separate the series into four components: trend, cycle, seasonality, and irregularity. Decomposition methods can be described globally as equation (4). The decomposition model can be used assuming no cyclical variation or the cyclical variation can be extracted and a model fitted so that to enhance future forecasting. Of course, there can be quite complex decomposition models, but the classical multiplicative decomposition model in NCSS was used here for this turnover series.

(4)

### Winters exponential smoothing (WES)

WES models work well on series that either have seasonality or seasonality and trend. The models can be additive or multiplicative as well, but the preferred option tends to have additive trend and either additive or multiplicative seasonality. Multiplicative trend in these kinds of time series models tends to over or under forecast for future values. Of course, dampened models can help avoid this future forecast problem, but one must be careful then that the forecasts are short term ([DeLurgio, 1998](#_ENREF_13)). The robustness and accuracy of Winter’s exponential smoothing methods has led to their widespread use in applications where a large number of series necessitates an automated procedure ([Winters, 1960](#_ENREF_40), [Taylor, 2003](#_ENREF_34)). WES model is easy to interpret and easily understood by management or those less technically inclined.

1. *Autoregressive Integrated Moving Average (ARIMA)*

The ARIMA models were introduced by the statisticians George Box and Gwilym Jenkins ([Box, Jenkins and Reinsel, 1970](#_ENREF_9)). The general form of ARIMA models is ARIMA(p,d,q); where *p*, *d*, and *q* are non-negative integers that refer to the order of the autoregressive, integrated, and moving average parts of the model respectively. In addition, ARIMA models can handle seasonality, and their forms would be as follows: ARIMA(p,d,q)(P,D,Q). Thus, the seasonal aspect of a series could have an autoregressive, differencing, or moving average patterns as well. ARIMA methods are popular to some forecasters because they provide a wide class of models for univariate time series forecasting ([Harvey and Todd, 1983](#_ENREF_19)).

### Univariate methods (with external variables)

Univariate models that incorporated an external variable (CLI) as a predictor of cyclical component of turnover series were also examined. Basically, this included dynamic regression and more complex decomposition models.

1. *Dynamic regression with external variable*

Dynamic regression model describes how the forecasting output is linearly related to current and past values of one or more input series. There are two crucial assumptions for dynamic regression model. First, the observations of the input series are assumed to occur at equally spaced time intervals. Second, the input series are not affected by the output ([Pankratz, 2012](#_ENREF_30)). Dynamic regression models allow one to include external variables, interventions, and transfer functions. In this study, the external variables (CLI and interventions) are incorporated into the dynamic regression model.

1. *Decomposition with external variable*

In this more complex decomposition model, CLI and its lag terms are used as a predictor of cyclical component in the decomposition model. This research applies two approaches to obtain a decomposition model: one is the decomposition built in NCSS with the cyclical variable (CLI) incorporated and the other is a multiplicative decomposition model built by the product of the best ARIMA and cyclical factor (CLI).

### Model evaluation

A good forecasting model should be evaluated on predictive ability, goodness of fit using the R2 value, mean absolute percentage error (MAPE), mean absolute error (MAE) or other fit diagnostics, normality tests on residuals, and a white noise test on those same residuals to make sure that no pattern is left. The best fitting model is generally selected with higher R2 value in the holdout data, lower MAPE, normally distributed residuals, and a passed white noise test.

1. *Pseudo R2*

For evaluation of the time series methods, the pseudo R2 for training and holdout data is calculated as standard criteria to test the goodness of model fitting taking the form in equation (5):

(5)

1. *Mean absolute percentage error*

MAPE is the other measure of accuracy of the time series model fitting methods as shown in equation (6) ([Hanke, Reitsch and Wichern, 1998](#_ENREF_18), [Bowerman, O'Connell and Koehler, 2005](#_ENREF_7)). This criterion is used to compare the model performance for the specific dataset by using time series methods, since it measures relative performance ([Chu, 1998](#_ENREF_11)).

(6)

1. *Normality test*

A good time series model should have normally distributed residuals. In this paper, normality of residuals is evaluated by two powerful normality tests: Shapiro-Wilk ([Shapiro and Wilk, 1964](#_ENREF_33)) and D'Agostino Omnibus normality test ([D'agostino, Belanger and D'Agostino Jr, 1990](#_ENREF_12)) in NCSS.

1. *White noise test*

The white noise test is performed on the residuals to evaluate whether there might be some time series pattern remaining in the dataset not accounted for the model and result in independent residuals, random scatter, or no more time series pattern in residuals ([Weisent, Seaver, Odoi and Rohrbach, 2010](#_ENREF_39)). In practice, the Q-statistic (also called Box-Pierce statistic or Ljung-Box statistic) is used as an objective diagnostic measure of white noise for a time series to compare whether the autocorrelations from residuals and white noise are statistically significantly different. This test statistic is illustrated in equation (7); where *k* is selected to be about the lesser of two seasonal cycles, about one-fourth of the observations, or 24 when two seasonal cycles is much greater than 24. In most cases, if a model is lacking in white noise, it means this model is deficient and has to be rectified ([DeLurgio, 1998](#_ENREF_13)).

(7)

# Results and Discussion

# *Temporal patterns of employ turnover and outlier identification*

The time based turnover series was restructured in a format so that it could be analysed to observe patterns like trend or seasonality and understand any inherent time series in the data for further investigation in that direction. As shown in Figure 1 (a), there is obvious seasonality pattern in the series. The box plot for turnover series from ANOVA test confirmed this seasonality pattern. In addition, there is a decreasing trend from January to November as shown in Figure 1 (b), and then the trend line rises up in December. The points with year label in Figure 1 (b) are considered as outliers, since they are beyond the upper whiskers. After all ARIMA models were tested, three outliers appeared in SPC chart (as shown in Figure 2); and two of these three were also flagged by the box plot analysis. Combining the outliers identified by ANOVA test and SPC, 7 outliers are smoothed to soften their impact. The outliers include the turnover numbers for December 2001, March 2008, April 2008, May 2008, June 2008, August 2008, and September 2008. Whenever one has outliers, one needs to investigate the cause. For instance, December 2001 is a 9/11 lag impact on job hiring with stronger background checking and increased retiring/hiring. The outliers in 2008 reflect the downsize policy in the organization, the uncertainty of the presidential election and spin-offs.

C:\Users\julia\Dropbox\Timeseries paper\figure\FIG1_TSBOXPLOT_FINAL_2.tif

Figure 1. (a) Monthly turnover series plotted over time and (b) box plot of turnover data.

Figure 2. SPC chart for standardized residuals.

The ACF and PACF plots for the turnover series are shown in Figure 3 (a, b). The pattern of unsmoothed data in the ACF and PACF hints at ARIMA(1,0,1) with some type of seasonality, but the seasonal pattern is not obvious.



Figure 3. (a) Autocorrelation and (b) partial autocorrelation plots.

# *Cross-correlations*

The cross-correlation is employed between turnover series and CLI series to identify significant lag correlations. The cross-correlations between first differences for turnover and the CLI series were examined ([DeLurgio, 1998](#_ENREF_13)), and a “pre-whitening” process for the two series was used to identify the cross-correlation patterns ([Box et al., 1970](#_ENREF_9), [Bowie and Prothero, 1981](#_ENREF_8), [Department of Sciences at Pennsylvania State University, 2014](#_ENREF_14)). Based all these calculation, CCFs from Lag 0 to Lag 8 are statistically significant, which indicates that the turnover series has statistically significant correlation with CLI and its 8 lags. The CLI and its 8 lags are applied into the dynamic regression, decomposition, and ARIMA model respectively as cyclical factor for forecasting.

# *Forecasting results and comparisons*

Forecasting evaluations for the univariate time series models are provided in the Appendix Table A. 1, Table A. 2, Table A. 3, and Table A. 4. Based on the evaluation statistics, 8 models are selected because of an acceptable R2 value for the training and holdout data as well as their residual statistics that are the optimum among the other models (as shown in Table 2). On average, the holdout R2 value of these models is 0.51 (range from 0.40 to 0.59).

### Univariate methods (without external variables)

The regression model with additive trend and seasonality has the highest holdout R2 (0.57) among the univariate models without external variables, indicating the model’s ability to explain 57% of the total variation of the holdout sample. It is statistically significant (P<0.05) for the model and coefficients. However, the residuals are not normally distributed and did not passed white noise test.

Table 2. Statistics for selected time series models.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Method | # | Model | Pred.1 R2 | Holdout R2 | MAPE | Normality2 | WN3 |
| Univariate without external variable | U1 | Regression with additive trend and seasonality | 0.51 | 0.57 | 26.15 | No | No |
| U2 | Regression with additive trend, seasonality and intervention4 | 0.72 | 0.52 | 22.84 | Yes | No |
| U3 | Decomposition | 0.65 | 0.54 | 17.97 | Yes | Yes |
| U4 | WES with additive trend and seasonality | 0.52 | 0.52 | 20.65 | Yes | Yes |
| U5 | ARIMA(1,0,1)(0,1,1) | 0.47 | 0.40 | 22.89 | Yes | Yes |
| Univariate with external variable | V1 | Dynamic regression using lag7 CLI as predictor4 | 0.77 | 0.59 | 19.91 | Yes | Yes |
| V2 | Decomposition using lag1 CLI as cycle | 0.65 | 0.55 | 17.97 | Yes | Yes |
| V3 | ARIMA combining with lag1 CLI as cycle | 0.37 | 0.41 | 22.73 | Yes | Yes |

Notes:

1. Pred. R2 is prediction R2 value for training data.
2. Normality is residuals’ normality test.
3. WN is white noise test.
4. The data is unsmoothed (or outliers are unadjusted) so as to take advantage of time series models that can accommodate interventions.

The regression model with additive trend, seasonality, and interventions (pulse and step) performs well with a training R2 of 0.72 and a holdout R2 of 0.52, indicating the model’s ability to explain 72% of the total variation of turnover for the training data and 52% of the total variation of the holdout sample. It is statistically significant (P<0.05) for the model and coefficients. The model has normal residuals, but it does not have a white noise residual pattern. However, this regression model can capture the spike in December 2001 and sharp fluctuations from March 2008 to August 2008 (as shown in the Figure 4).

Figure 4. Regression with interventions forecast vs. actual turnover number plot.

The decomposition model is considered as the best univariate model without external variables because this model has a reasonably high training R2 value (0.65), a good holdout R2 value (0.54), and low MAPE (17.97). The residuals of this model are normally distributed and have a white noise pattern. Figure 5 shows the predicted turnover based on the decomposition model and the actual turnover number for holdout dataset. This plot validates the holdout performance of the decomposition model as it is able to mimic the changes in trend and seasonality of the turnover and prediction is close to the actual turnover numbers. However, it does seem to under-forecast for the 6 months June through November.

Figure 5. Decomposition validation forecast vs. actual turnover number with prediction intervals.

### Univariate methods (with external variables)

According to the cross-correlation analysis result, CLI and its 8 lags are applied in dynamic regression, decomposition model, and ARIMA(1,0,1)(0,1,1) respectively as external variable to forecast turnover number. The dynamic regression model with additive trend, seasonality, interventions (pulse and step), and lag7 of CLI is the best model among all models, since it has highest predicted and holdout R2 value (0.77 and 0.59), normalized residuals, and white noise. The dynamic regression is globally statistically significant and individually significant for the parameter coefficients (P<0.05). Figure 6 shows the predicted and actual turnover number plots for holdout dataset from dynamic regression model. Although there are under forecasts from July to September as well, the differences between forecasting value and actual value have become much tighter. Compared with top rated univariate methods without external variables, the performance of dynamic regression model is much improved after using CLI as cyclical factor.

Figure 6. Dynamic regression with lag7 CLI forecast vs. actual turnover number for holdout dataset.

# Conclusion

In this paper, various univariate time series forecasting models for employee turnover prediction are tested and optimal models for turnover forecasts are identified. The model in the paper actually performs better than those accessed in the literature review. Dynamic regression model is concluded as the best forecasting model. Dynamic regression model has several advantages. For example, ARIMA error term which has autocorrelation pattern can be included in the model. Dynamic regression model is able to handle lagged regressors and various types of seasonality. In addition, dynamic regression model can handle interventions or change points effectively, since these interventions or change points, such as holidays, promotions, new policy and so forth, are often common to the time series data. Thus, dynamic regression model could be used to forecast turnover for most of organizations including small size, large size, and merged unit. However, to do dynamic regression modelling, at least 5-years of monthly employee turnover data is preferred to make accurate forecast. If the horizon of dataset is less than 5 years, a special decomposition model ([Ittig, 1997](#_ENREF_23)) could be considered as a substitute. Even though the forecasting horizon provided by dynamic regression model is relatively short, this is not a big issue for human resource departments, since most of human resource departments are only interested in a short term, such as three months, forecast. Therefore, if the human resource department in an organization is not well familiar with forecasting techniques, dynamic regression model could be a good option for a preliminary turnover forecast once a CLI can be developed as in this paper. It is worth mentioning that external variable like CLI in this paper does help on the turnover forecast. Especially, when the human resource department has a small and unreliable dataset, incorporating the external variables such as CLI may help the whole forecasting process. If the external variable such as CLI is not available, decomposition model could be considered as the first choice rather than dynamic regression model. In practice, some software such as NCSS has an embedded decomposition macro, which the human resource departments are able to run the decomposition model quite easily.

# *Practical implications*

1. According to our findings, employee turnover forecast, in practice, could be handled easily. We suggest that human resource departments could use regression model for the preliminary forecast and the accuracy of forecast is acceptable. However, there are some types of interventions that regression cannot handle, such as pulse or steps with exponential decay or growth. Dynamic regression is likely to be used as an alternative for forecasting.
2. In this study, statistical analysis packages SAS and NCSS are used for the forecast. However, to some organizations, the human resource departments are unwilling to spend extra funding on software purchase. Under this circumstance, Microsoft Excel could be a good alternative for the time series forecast without extra investment, because there have been some open source time series forecasting packages designed to run under Excel environment in the market. The forecasting models such as naïve model, moving average, exponential smoothing, decomposition, regression, and ARIMA model have been included into these packages ([Warren, 2008](#_ENREF_38)).
3. Another option these days is to use R for the dynamic regression ([Hyndman, 2014](#_ENREF_22)). The pseudo R code for dynamic regression and test on residuals (normality and white noise test) is provided below.

## Install R package

library (dynlm)

library (normwhn.test)

## “Turnover” is turnover data.

## “Trend” is a continuous variable with value from 1 to n.

## “Seasonality” is dummy variable representing seasonal periods.

## “X” is the lag term of CLI.

## “Intervention” is external variable impact: yes=1and no=0.

## Build model ## Load dataset

Turnover = read.xls (“turnover\_data\_in\_excel.xls”)

## IMPORTANT: Some R codes in dynamic regression may not handle fancy intervention analysis.

Turnover.Model = dynlm (Turnover ~ Trend + Seasonality + L(CLI, X) + Intervention)

## Model summary

summary ( Turnover.Model)

## Calculate residuals

Residual = resid (Turnover.Model)

## Normality test on residuals

normality.test1 (Residual)

## White noise test on residuals

whitenoise.test (Residual)

# *Limitations and future research*

This study is limited to forecast total turnovers in an organization. It may be possible to apply univariate time series forecasting models to forecast turnovers in different categories like retirement or volunteer quit. Although the study incorporated CLI as an external factor and the accuracy of forecasts is well improved, the external factors affecting turnover can be much beyond the scope of CLI, as local and cyclical economic fluctuations strongly influence the propensity of employees to quit ([Abelson and Baysinger, 1984](#_ENREF_1)).

# Reference

Abelson, M.A. and Baysinger, B.D. (1984), 'Optimal and Dysfunctional Turnover: Toward an Organizational Level Model', *Academy of Management Review*, 9, 331-341.

Alao, D. and Adeyemo, A. (2013), 'Analyzing Employee Attrition Using Decision Tree Algorithms ', *Computing, Information Systems, Development Informatics and Allied Research Journal*, 4, 17-28.

Alwan, L.C. and Roberts, H.V. (1988), 'Time-series Modeling for Statistical Process Control', *Journal of Business & Economic Statistics*, 6, 87-95.

Balfour, D.L. and Neff, D.M. (1993), 'Predicting and Managing Turnover in Human Service Agencies: A Case-study of an Organization in Crisis ', *Public Personnel Management*, 22, 473-486.

Bergmeir, C. and Benítez, J.M. (2012), 'On the Use of Cross-validation for Time Series Predictor Evaluation', *Information Sciences*, 191, 192-213.

Bluedorn, A.C. (1982), 'A Unified Model of Turnover from Organizations', *Human Relations*, 35, 135-153.

Bowerman, B.L., O'Connell, R.T. and Koehler, A.B. (2005), *Forecasting, Time Series, and Regression: An Applied Approach*: South-Western Pub.

Bowie, C. and Prothero, D. (1981), 'Finding Causes of Seasonal Diseases Using Time Series Analysis', *International journal of epidemiology*, 10, 87-92.

Box, G.E., Jenkins, G.M. and Reinsel, G.C. (1970), *Time Series Analysis: Forecasting and Control*: San Francisco: Holden Day.

Čambál, M., Holková, A. and Lenhardtová, Z. (2011), 'Basics of the Management': Trnava: AlumniPress.

Chu, F.-L. (1998), 'Forecasting Tourist Arrivals: Nonlinear Sine Wave or ARIMA?', *Journal of Travel Research*, 36, 79-84.

D'agostino, R.B., Belanger, A. and D'Agostino Jr, R.B. (1990), 'A Suggestion for Using Powerful and Informative Tests of Normality', *The American Statistician*, 44, 316-321.

DeLurgio, S.A. (1998), *Forecasting Principles and Applications*: McGraw-Hill/Irwin, New York.

Pre-whitening as an Aid to Interpreting the CCF. https://onlinecourses.science.psu.edu/stat510/?q=node/75.

Feeley, T.H. and Barnett, G.A. (1997), 'Predicting Employee Turnover From Communication Networks', *Human Communication Research*, 23, 370-387.

Größler, A. and Zock, A. (2010), 'Supporting Long-term Workforce Planning with a Dynamic Aging Chain Model: A Case Study from the Service Industry', *Human Resource Management*, 49, 829-848.

Grznar, J., Booth, D.E. and Sebastian, P. (1997), 'A Robust Smoothing Approach to Statistical Process Control', *Journal of Chemical Information and Computer Sciences*, 37, 241-248.

Hanke, J.E., Reitsch, A.G. and Wichern, D.W. (1998), *Business Forecasting*: Prentice Hall Upper Saddle River, NJ.

Harvey, A.C. and Todd, P. (1983), 'Forecasting Economic Time Series with Structural and Box-Jenkins Models: A Case Study', *Journal of Business & Economic Statistics*, 1, 299-307.

Heneman, H.G., Schwab, D.P., Fossum, J.A. and Dyer, L.D. (1993), *Personnel/Human Resource Management*: Irwin Homewood, IL.

Hong, W.-C., Wei, S.-Y. and Chen, Y.-F. (2007), 'A Comparative Test of Two Employee Turnover Prediction Models', *International Journal of Management*, 24, 216-229.

CRAN Task View: Time Series Analysis. <http://cran.r-project.org/web/views/TimeSeries.html>.

Ittig, P.T. (1997), 'A Seasonal Index for Business', *Decision Sciences*, 28, 335-355.

Khoong, C. (1996), 'An Integrated System Framework and Analysis Methodology for Manpower Planning', *International Journal of Manpower*, 17, 26-46.

Leonard, B. (2001), 'Turnover at the Top', *HR Magazine*, 46-52.

Morrow, P.C., McElroy, J.C., Laczniak, K.S. and Fenton, J.B. (1999), 'Using Absenteeism and Performance to Predict Employee Turnover: Early Detection through Company Records', *Journal of Vocational Behavior*, 55, 358-374.

Nagadevara, V., Srinivasan, V. and Valk, R. (2008), 'Establishing a Link between Employee Turnover and Withdrawal Behaviours: Application of Data Mining Techniques ', *Research & Practice in Human Resource Management*, 16, 81-99.

Ng, S.H., Cram, F. and Jenkins, L. (1991), 'A Proportional Hazards Regression Analysis of Employee Turnover among Nurses in New-Zealand ', *Human Relations*, 44, 1313-1330.

Composite Leading Indicators(MEI). <http://stats.oecd.org/index.aspx?queryid=6617>.

Pankratz, A. (2012), *Forecasting with Dynamic Regression Models*: John Wiley & Sons.

Saradhi, V.V. and Palshikar, G.K. (2011), 'Employee Churn Prediction', *Expert Systems with Applications*, 38, 1999-2006.

Sexton, R.S., McMurtrey, S., Michalopoulos, J.O. and Smith, A.M. (2005), 'Employee Turnover: A Neural Network Solution', *Computers & Operations Research*, 32, 2635-2651.

Shapiro, S.S. and Wilk, M.B. (1964), 'An Analysis of Variance Test for Normality (Complete Samples)', JSTOR.

Taylor, J.W. (2003), 'Short-term Electricity Demand Forecasting Using Double Seasonal Exponential Smoothing', *Journal of the Operational Research Society*, 54, 799-805.

Thaden, E., Jacobs-Priebe, L. and Evans, S. (2010), 'Understanding Attrition and Predicting Employment Durations of Former Staff in a Public Social Service Organization', *Journal of Social Work*, 10, 407-435.

Velicer, W.F. and Fava, J.L. (2003), 'Time Series Analysis', *Handbook of psychology*.

Velleman, P.F. (1980), 'Definition and Comparison of Robust Nonlinear Data Smoothing Algorithms', *Journal of the American Statistical Association*, 75, 609-615.

Forecasting Time Series within Excel. <http://web.calstatela.edu/faculty/hwarren/a503/forecast%20time%20series%20within%20Excel.htm>.

Weisent, J., Seaver, W., Odoi, A. and Rohrbach, B. (2010), 'Comparison of Three Time-series Models for Predicting Campylobacteriosis Risk', *Epidemiology and infection*, 138, 898-906.

Winters, P.R. (1960), 'Forecasting Sales by Exponentially Weighted Moving Averages', *Management Science*, 6, 324-342.

Wright, T.A. and Cropanzano, R. (1998), 'Emotional Exhaustion as a Predictor of Job Performance and Voluntary Turnover', *Journal of Applied Psychology*, 83, 486-493.

Zhang, G.P. (2003), 'Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model', *Neurocomputing*, 50, 159-175.

# Appendix

Table A. 1. Time series univariate models for turnover data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Variables | Pred. R2 | Holdout R2 | MAPE | Normality | WN |
| Regression1 | Additive T+S | 0.51 | 0.57 | 26.15 | No | No |
| Additive S | 0.45 | 0.09 | 22.32 | No | No |
| Additive T+S+T\*S | 0.58 | 0.22 | 23.41 | No | No |
| Additive T+S+Intervention2 | 0.72 | 0.52 | 22.84 | Yes | No |
| Multiplicative T.S | 0.34 | 0.55 | 25.42 | Yes | No |
| Multiplicative S | 0.33 | 0.50 | 25.58 | Yes | No |
| Multiplicative T.S (T\*S) | 0.30 | 0.26 | 29.75 | Yes | No |
| Decomposition1 | T\*C\*S | 0.65 | 0.54 | 17.97 | Yes | Yes |
| WES1 | Additive T+ Additive S | 0.52 | 0.52 | 20.65 | Yes | Yes |
| Additive T+ Multiplicative S | 0.54 | 0.46 | 20.17 | Yes | Yes |
| Multiplicative T+ Additive S | 0.52 | 0.40 | 19.82 | Yes | Yes |
| Multiplicative T+ Multiplicative S | 0.52 | 0.39 | 19.77 | Yes | Yes |
| ARIMA | ARIMA(1,0,0)(2,1,0) | 0.39 | 0.33 | 22.94 | No | Yes |
| ARIMA(1,0,0)(2,0,0) | 0.38 | 0.38 | 23.96 | Yes | Yes |
| ARIMA(0,1,1)(2,1,0) | 0.45 | 0.14 | 22.96 | Yes | Yes |
| ARIMA(0,1,1)(2,0,1) | 0.42 | 0.17 | 22.45 | Yes | Yes |
| ARIMA(0,0,1)(1,0,1) | 0.35 | 0.30 | 25.35 | Yes | Yes |
| ARIMA(0,0,1)(1,0,2) | 0.39 | 0.33 | 22.96 | Yes | No |
| ARIMA(1,0,0)(0,1,1) | 0.40 | 0.48 | 25.20 | Yes | Yes |
| ARIMA(1,0,1)(2,0,0) | 0.43 | 0.26 | 21.82 | Yes | Yes |
| ARIMA(1,0,1)(1,0,2) | 0.44 | 0.20 | 21.55 | Yes | Yes |
| ARIMA(1,0,1)(0,1,1) | 0.47 | 0.40 | 22.89 | Yes | Yes |
| ARIMA(2,1,0)(0,1,1) | 0.41 | 0.26 | 25.24 | Yes | Yes |
| ARIMA(0,0,1)(2,0,0) | 0.36 | 0.37 | 25.00 | Yes | Yes |
| ARIMA(1,1,1)(1,0,0) | 0.35 | 0.04 | 24.20 | No | Yes |
| ARIMA(1,0,1)(1,1,0) | 0.39 | 0.32 | 22.45 | Yes | Yes |
| ARIMA(1,0,1)(1,0,1) | 0.41 | 0.24 | 22.18 | No | Yes |
| ARIMA(0,1,1)(2,0,0) | 0.40 | 0.16 | 22.60 | Yes | Yes |
| ARIMA(1,1,1)(2,0,0) | 0.41 | 0.20 | 22.26 | Yes | Yes |
| ARIMA(1,0,1)(1,0,1) | 0.35 | 0.04 | 23.87 | No | Yes |

Note: 1. In Regression, Decomposition and Exponential Smoothing models: T denotes Trend, S denotes seasonality, C denotes cycle, and T\*S denotes an interaction term between T and S.

2. The data is unsmoothed (or outliers are unadjusted) so as to take advantage of time series models that can accommodate interventions.

Table A. 2. Dynamic regression model with CLI and its lags as cyclical factor.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Pred. R2 | Holdout R2 | MAPE | Normality | WN |
| Dynamic regression & CLI | 0.72 | 0.56 | 22.56 | Yes | No |
| Dynamic regression & lag1 CLI | 0.72 | 0.54 | 22.73 | Yes | No |
| Dynamic regression & lag2 CLI | 0.74 | 0.55 | 21.87 | Yes | No |
| Dynamic regression & lag3 CLI | 0.74 | 0.55 | 21.50 | Yes | Yes |
| Dynamic regression & lag4 CLI | 0.74 | 0.56 | 21.35 | Yes | Yes |
| Dynamic regression & lag5 CLI | 0.75 | 0.57 | 21.32 | Yes | Yes |
| Dynamic regression & lag6 CLI | 0.76 | 0.58 | 20.73 | Yes | Yes |
| Dynamic regression & lag7 CLI | 0.77 | 0.59 | 19.91 | Yes | Yes |
| Dynamic regression & lag8 CLI | 0.77 | 0.59 | 20.05 | Yes | Yes |

Table A. 3. Decomposition model with CLI and its lags as cyclical factor.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Pred. R2 | Holdout R2 | MAPE | Normality | WN |
| Decomposition & CLI | 0.65 | 0.55 | 17.97 | Yes | Yes |
| Decomposition & lag1 CLI | 0.65 | 0.55 | 17.97 | Yes | Yes |
| Decomposition & lag2 CLI | 0.65 | 0.55 | 17.97 | Yes | Yes |
| Decomposition & lag3 CLI | 0.65 | 0.55 | 17.97 | Yes | Yes |
| Decomposition & lag4 CLI | 0.65 | 0.54 | 17.97 | Yes | Yes |
| Decomposition & lag5 CLI | 0.65 | 0.54 | 17.97 | Yes | Yes |
| Decomposition & lag6 CLI | 0.65 | 0.53 | 17.97 | Yes | Yes |
| Decomposition & lag7 CLI | 0.65 | 0.53 | 17.97 | Yes | Yes |
| Decomposition & lag8 CLI | 0.65 | 0.53 | 17.97 | Yes | Yes |

Table A. . ARIMA model with CLI and its lags as cyclical factor.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Pred. R2 | Holdout R2 | MAPE | Normality | WN |
| ARIMA(1,0,1)(0,1,1) & CLI | 0.37 | 0.41 | 22.77 | Yes | Yes |
| ARIMA(1,0,1)(0,1,1) & lag1 CLI | 0.37 | 0.41 | 22.73 | Yes | Yes |
| ARIMA(1,0,1)(0,1,1) & lag2 CLI | 0.37 | 0.41 | 22.78 | Yes | Yes |
| ARIMA(1,0,1)(0,1,1) & lag3 CLI | 0.37 | 0.41 | 22.85 | Yes | Yes |
| ARIMA(1,0,1)(0,1,1) & lag4 CLI | 0.36 | 0.40 | 22.91 | Yes | Yes |
| ARIMA(1,0,1)(0,1,1) & lag5 CLI | 0.36 | 0.40 | 22.98 | Yes | Yes |
| ARIMA(1,0,1)(0,1,1) & lag6 CLI | 0.36 | 0.40 | 23.04 | Yes | Yes |
| ARIMA(1,0,1)(0,1,1) & lag7 CLI | 0.36 | 0.40 | 23.09 | Yes | Yes |
| ARIMA(1,0,1)(0,1,1) & lag8 CLI | 0.36 | 0.40 | 23.12 | Yes | Yes |

1. \* Corresponding author: Shuguang Ji. Phone: +1.865.951.8083. Email: [sji1@utk.edu](mailto:sji1@utk.edu). [↑](#footnote-ref-1)
2. \* Corresponding author. Email: [sji1@utk.edu](mailto:sji1@utk.edu). [↑](#footnote-ref-2)