The Report Committee for Srikanth Tondukulam Seetharaman
Certifies that this is the approved version of the following report:

Forecasting of sick leave usage among nurses via artificial neural networks

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

John J Hasenbein

Elmira Popova

Forecasting of sick leave usage among nurses via artificial neural networks

by

Srikanth Tondukulam Seetharaman, B.Tech.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

December 2010

Dedication

To *my parents Padma and Seetharaman and brother Kumar,*

Who have always been the greatest source of motivation for me

To my grandparents *Kalyani Krishnan, Late Krishnan and Late Rajalakshmi,*

Whose blessings have always been with me

To my relatives *Anantha Narayanan, Savithri and Vijayalakshmi,*

Who have continuously supported me in all my endeavors

To my professors *Dr A.K.Bakthavatsalam and Mr R. Gururaj,*

Who have always inspired me to discover my potential

Acknowledgements

I would like to thank my advisor, Dr. John J Hasenbein for granting me the opportunity to work with him and guiding me through the course of this report. I would also like to acknowledge Dr. Elmira Popova for her participation. Thanks to the faculty and staff at the ORIE department of UT Austin for contributing to my education.

I further extend my special thanks to Dr. Anura deSilva, Dr. Lalith deSilva, Andres Granados, Shankar Dhanaraj, Aniruddha Kulkarni and the whole team at Planmatics, Inc. for their support.

I would like to highlight the enormous role that my parents and brother have played in encouraging me. A lot of credit also goes to my relatives who have supported me. I am lucky to be gifted with valuable friends who have inspired me at every stage of my life. I would like to thank my apartment mates in Austin, Rahul and Sambuddha, for all the great times we have shared. Thanks also to Daniel Moore, Vaidy, Naveed, Josh Adams, John Cherukuri, Vijo Varkey, Jagannath, Sundeep, Vamsi and Onesi for making my stint in Austin all the more memorable.

I would like to wind this up by thanking some friends who are very close to me. This includes the Cicreps ’06, my colleagues at L&T, my prod family of Dev, Vigi and GK and finally, the closest of all, the KoolKandus.

11/23/2010

Abstract

Forecasting of sick leave usage among nurses via artificial neural networks

Srikanth Tondukulam Seetharaman, M.S.E.

The University of Texas at Austin, 2010

Supervisor: John J Hasenbein

This report examines the trends in sick leave usage among nurses in a hospital and aims at creating a forecasting model to predict sick leave usage on a weekly basis using the concept of artificial neural networks (ANN). The data used for the research includes the absenteeism (sick leave) reports for 3 years at a hospital. The analysis shows that there are certain factors that lead to a rise or fall in the weekly sick leave usage. The ANN model tries to capture the effect of these factors and forecasts the sick leave usage for a 1 year horizon based on what it has learned from the behavior of the historical data from the previous 2 years. The various parameters of the model are determined and the model is constructed and tested for its forecasting ability.

Table of Contents

List of Tables viii
List of Figures ix
1. Introduction 1
    1.1 Artificial neural networks 2
        1.1.1 Concept and architecture 2
        1.1.2 Main parameters for construction 4
        1.1.3 Performance measures 5
    1.2 ANN forecasting 6
        1.2.1 Training cycle 6
        1.2.2 Cross validation cycle 7
        1.2.3 Forecasting cycle 8
2. Literature review 9
3. Analysis 12
    3.1 Definitions 12
    3.2 Holiday analysis 13
4. Proposed model 22
    4.1 ANN model construction 22
    4.2 Building the input matrix 23
    4.3 ANN behavior 26
5. Results 29
6. Conclusion 32
7. References 33
Vita 35

List of Tables

Table 1: Definition of weeks and years 12
Table 2: Holiday calendar 14
Table 3: Expansion for column names used in input matrix 25
Table 4: Data to study ANN behavior 27
Table 5: Iteration checkpoints 31
Table 6: Experiment for testing consistency of forecast results 31

List of Figures

Figure 1: Basic ANN layers 3
Figure 2: ANN model structure 4
Figure 3: Training cycle 6
Figure 4: Forecasting cycle 8
Figure 5: Weekly sick leave usage 13
Figure 6: New Year and Epiphany effects 16
Figure 7: Carnival effects 16
Figure 8: Easter effects 17
Figure 9: May Day effects 18
Figure 10: Ascension Day, Whit week and Corpus Christi effects 19
Figure 11: Unity, All Saints, Repentance and Fall vacations effects 19
Figure 12: Christmas effects 20
Figure 13: Summer vacations effects 21
Figure 14: Input matrix for year 2007 24
Figure 15: Forecast results for cases A and B 28
Figure 16: Training cycle progress 29
Figure 17: Actual and forecasted outputs for year 2008 30

1. Introduction

The job of a nurse is physically and mentally challenging. Nurses are exposed to illness daily and their job is inherently stressful. Hence sick leaves are an important factor in a nurse's job routine and their usage or abuse makes for interesting material for analysis.

In June 2005, the UK Guardian reported that nurses not only top the public sector sick leave table, but also record 50% more sick leaves than other public sector workers. The study revealed that nurses and healthcare assistants take, on average, 16.8 days a year of sick leave. These figures look ominous when compared with the 11.3 days across other sectors of the public workforce, including teachers, police officers, social workers and civil servants. The scenario is not too different in other countries, including the US.

One of the greatest motivations to call in sick when you are actually fit would be to create extended vacations. Hence, there is a tendency to call in sick during shifts that fall close to or in the same week as holidays. The usage of sick leaves can also be seasonal, with the numbers rising with the advent of the flu season. In workplaces where the leave policies allow employees to take a fixed number of leaves per year, the rate of calling in sick could increase towards the end of the cycle as employees try desperately to use up their quota of leaves.

The above points are just intuitively stated and cannot be confirmed without a proper statistical analysis of the sick leave usage. Sick leaves and other such leaves contribute to a category of unexpected absences which lead to a staffing problem. When a nurse calls in sick, he or she can be replaced in several ways. The workforce manager uses the most cost effective option which does not hamper the quality of service provided

as far as possible. The options include use of floating pool nurses, agency nurses, on-call nurses or overtime for registered nurses. But usually, all of these options increase expenses thus driving the need for preemptive action to deal with such circumstances. If sick leave usage can be predicted to the best accuracy possible and taken into account while doing the staffing beforehand, a lot of extra expenses can be averted.

There are many advanced techniques that are used for forecasting these days. This report focuses on the use of artificial neural networks (ANNs). This is a data driven technique which is applicable to a data intensive environment like the one in question. It was decided to experiment with the performance of ANN on this problem.

1.1 Artificial neural networks

1.1.1 Concept and architecture

The following definition of artificial neural networks is from Wikipedia:

> An artificial neural network (ANN), usually called "neural network" (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase ("Artificial neural network," 2010, para. 1)

We can see that ANNs are inspired by biological neural networks. The model uses a simple feed-forward multilayer architecture. A model consists of 3 main layers, namely the input layer, the transfer function and the output layer. The basic layers of ANNs are shown in Figure 1. Each of these layers contains neurons or nodes which form the building blocks of an ANN network. The network is completed by arcs which connect the nodes between different layers.

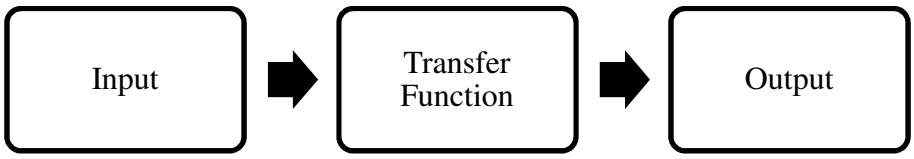

Figure 1: Basic ANN layers

A more detailed construction of ANN networks is depicted in Figure 2. The input layer contains nodes which accept the inputs of the equation. These inputs include the fields or factors that determine the nature or value of the output. An input layer can have any number of nodes.

The transfer function layer strictly serves as a mathematical computation layer which produces functions to convert inputs to outputs. In other words, it defines a relationship between the input and the output layers. This layer is made up one or more hidden layers which in turn contain nodes. The arcs that connect the nodes of these hidden layers to each other as well as to the input and the output layers are associated with certain parameters called arc weights. Arc weights are adjustable parameters of the transfer functions that cause the mathematical conversions at every stage of data flow

from input nodes through to the output nodes. The nodes in the hidden layer can be considered as hosts for intermediate calculation values.

The output layer contains the nodes that bear the value of the field to be forecasted.

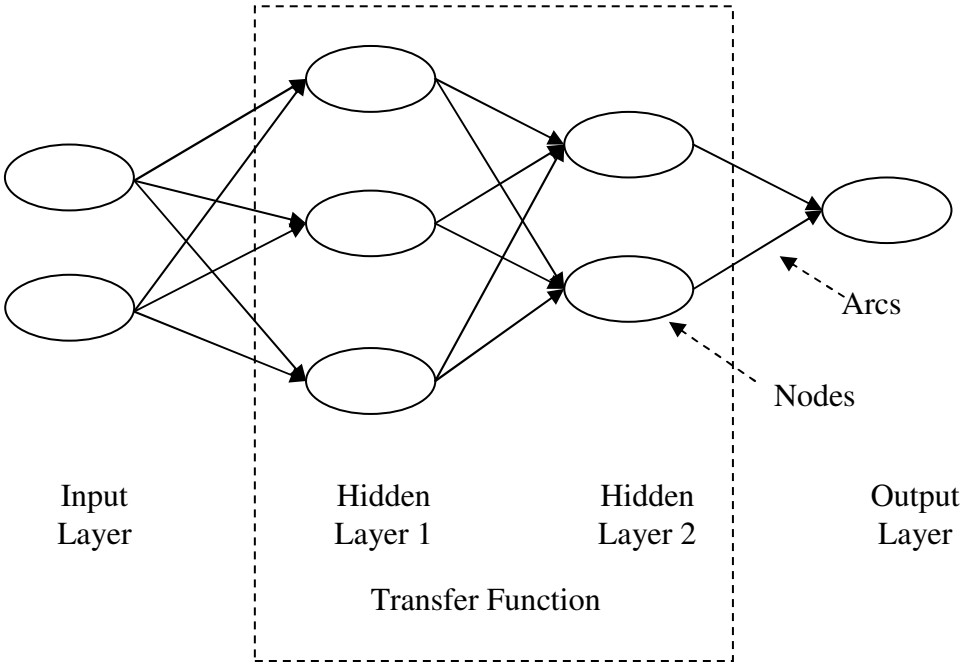

Figure 2: ANN model structure

1.1.2 Main parameters for construction

Six parameters have been identified to be critical to the performance of an ANN model. These are listed below

1. Number of input nodes (columns) - Every field that helps capture the behavior of the output counts as an input node.
2. Number of output nodes (column) - The output nodes are simply the fields that have to be forecasted.

3. Number of data points (rows/weeks) - A data point can be seen as a vector where the input and output nodes are the elements. Hence, every data point has a complete set of input and output node values assigned to it exclusively.
4. Number of hidden layers - These constitute the transfer function layer. A model can have any number of hidden layers.
5. Number of nodes in hidden layers - These are the nodes present in the hidden layers.
6. Number of iterations - A training cycle or run consists of one or more iterations. This parameter decides the performance of a model once the structure of the model is finalized. The role of iterations are explained in detail in the ANN forecasting section.

While parameters 1, 2 and 3 are usually determined during the first stage of the model construction, the values for parameters 4, 5 and 6 are arrived upon by trial and error methods after the first 3 parameters have been decided.

1.1.3 Performance measures

There are different ways of estimating the performance of the model. Some of the techniques that ANN commonly uses are mean squared error (MSE), mean absolute percentage error (MAPE) and the normalized root mean squared error (NMSE). A model can use one or a combination of the various measures. The main purpose of defining the performance measure is to quantify the performance of the model so that the model can work towards improving the performance measure.

1.2 ANN forecasting

Forecasting using ANN is done in 3 stages and each stage can be called a cycle.

1.2.1 Training cycle

The first stage, called the training cycle, is where the model is trained using historical data. During this cycle, all historical data points that are chosen and tagged as training data are fed into the model. The flow of data during this cycle is shown in Figure 3.

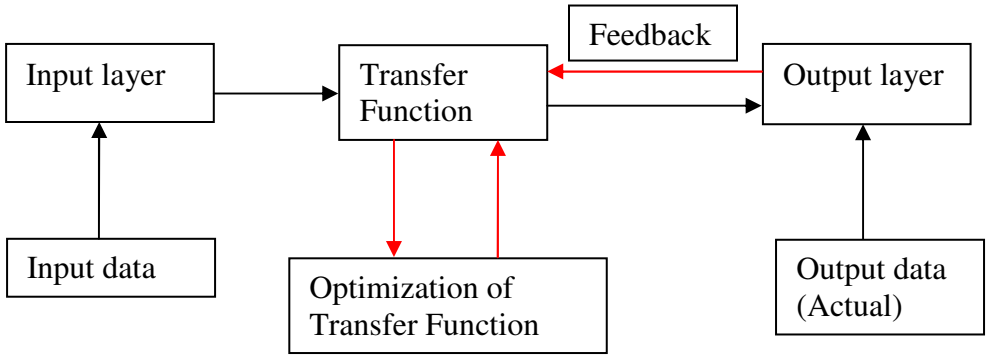

Figure 3: Training cycle

The values of the input nodes of all the data points are fed into the input layer. This data passes through the hidden nodes in the transfer function layer where the input is mathematically translated to come up with the output. This output that has been produced is compared with the actual values of the output nodes which are fed into the output layer. The error between the actual and the calculated output is determined and sent as a feedback to the transfer function layer. The transfer function layer in turn modifies the arc weights to try and achieve a better performance measure value. This can also be called optimization of the transfer function. If a model uses MSE as the performance measure, this means that for every iteration in the training cycle, the mean

squared error between the calculated and the actual outputs is determined and the goal of optimization is to minimize the mean squared error.

This completes one iteration in the training cycle. The input data is now treated with the new optimized transfer function and a new output is produced which is compared with the actual output values again in this new iteration. Every time, the mean squared error is improved by optimizing the transfer function which in turn is done by altering the arc weights.

1.2.2 Cross validation cycle

The next stage is called the cross validation cycle. This is not a mandatory stage and in our forecasting model, we have decided to skip this stage. This cycle runs in tandem with the training cycle, which means that every iteration in the training cycle would be immediately followed by an iteration in the cross validation cycle before moving to the next training iteration. Usually, a portion of the training data, around 10-20% is used as cross validation data. This means that this data is not used in training, but the performance of the model is tested on the data that is tagged as cross validation data by running a mock forecast on the cross validation data. We can opt to optimize the model based on the feedback from the training cycle, cross validation cycle or both. If the model is designed to use feedback from the cross validation cycle only, then a performance measure is calculated at the end of the mock forecast by comparing the forecasted outputs with the actual outputs during each iteration. The model is then trained to improve on this performance measure value in the next iteration.

1.2.3 Forecasting cycle

The final stage is the actual forecasting cycle, where the data points fed into the model are inputs with yet unknown outputs. These could represent days or weeks or years in the future in a timeline for which the prediction is being made. Figure 4 summarizes the behavior of the model during this cycle.

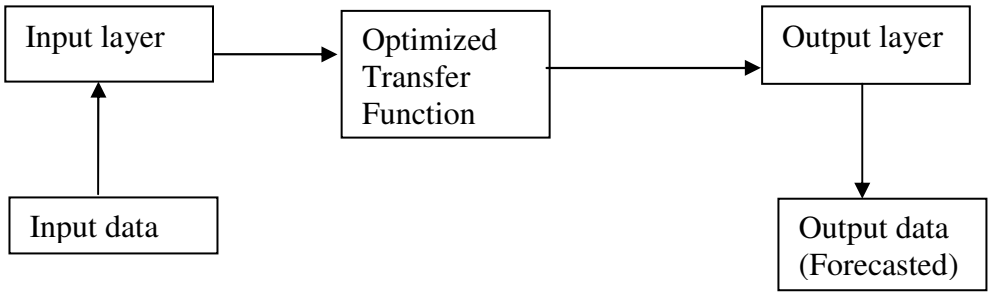

Figure 4: Forecasting cycle

The input data is fed into the input layer. These pass through the transfer function layer where the final optimized transfer function obtained from the training stage/ cross validation stage has been loaded. This optimized transfer function contains the best arc weights, which are also usually the weights from the last iteration. The input is transformed into a forecasted output which is produced from the model as the final result. If we are just testing the model for its forecasting ability versus doing some real time forecasting, we could provide the known actual outputs for the inputs of the forecasting cycle. The forecasting cycle performs the forecasting, plainly calculates the error in forecasting and displays it. The model can be instructed to display the mean squared error, maximum MSE, Minimum MSE, Normalized mean squared error, etc., in forecasting.

2. Literature review

This section reviews a portion the literature on forecasting using artificial neural networks. For starters, some papers on sick leave and absenteeism analysis were reviewed, but the bulk of the work done in this area deals with the study of the causes and human factors behind high sick leave usage. Hacket et al. (1989) perceived absenteeism as a volitional behavior and performed an idiographic-longitudinal analysis of absenteeism among hospital nurses. They looked into the exact reasons why nurses call in sick or take a day off. Clearly, their research aims at solving the sick leave abuse problem though a sociological approach. But since we are interested in forecasting the sick leave usage rather than trying to prevent abuse, our literature review is almost completely concentrated on ANN and its forecasting applications. There are many papers which compare the results produced by ANN based forecasting models to those produced by other approaches. These papers were browsed to understand the basic concepts of ANN and performance of ANN based forecasting models.

According to Zhang et al. (1998), ANN makes a good forecasting technique because it is a data driven approach with a self-learning and generalizing capability. Irie and Miyake (1988) have shown how a three layered model is necessary and sufficient to represent arbitrary functions. Hornik et al. (1989) displays the approximation capabilities of multi-layer feedforward networks in his works and follows it up with some more promising results (Hornik 1991, 1993).

The construction of ANN models has been dealt with in many previous works. Deciding the number of input and output nodes seems like a relatively easy task. Zhang

et al. (1998) state that ideally we must have a small number of essential nodes which can unveil the unique features embedded in the data. Tang et al. (1991) and Lachtermacher and Fuller (1995) have experimented with varying the number of input and output nodes but it ultimately comes down to which fields the researcher identifies as inputs and outputs when he carefully considers the problem and analyzes the data available.

The quest to find an ideal approach to decide on the number of hidden layers and hidden nodes has attracted a lot of research interest. Zhang et al. (1998) conclude from their research that for most problems, a single hidden layer is sufficient but that such a configuration may create the need for more hidden nodes, which is not desirable. Srinivasan et al. (1994) and Zhang (1994) show that two hidden layers result in a higher efficiency in training and a higher accuracy in forecasting. There appears to be a consensus that more than two hidden layers fail to add any significant forecasting capacity to the model.

Zhang et al. (1998) state that having too many hidden nodes does not do any good for the generalization characteristics of the model, leading to overfitting problems while having too few hidden nodes may affect the model's ability to learn the data accurately. However, there are certain cases (Lippmann 1987; Tang and Fishwick 1993; Kang 1991) where standard methods of allotting hidden nodes have been recommended. The ideal number of hidden nodes is usually recommended to be between 1 and $(2n + 1)$ where n is the number of input nodes. Tang and Fishwick (1993) maintain that the number of hidden nodes does not have any significant impact on the performance of the model.

ANN has widely been used as a forecasting tool. Al Saba and El Amin (1999) applied ANN techniques to long term load forecasting and found that the ANN results were closer to the actual data when compared to time series models. Chen et al. (2001) identified pricing as the determining factor and hence used this quantity in an input node for electricity demand in their ANN based short term load forecasting model. A variety of performance measuring tactics have been used previously and Chakraborty et al. (1992), Kang (1991) and Kohzadi et al. (1996) have successfully demonstrated the use of mean squared errors as an effective performance measure.

3. Analysis

The data analysis described below is necessary to identify the input nodes, output nodes and data points for the forecasting model.

3.1 Definitions

Based on the leave database pulled out from the scheduling system of a hospital, the start and end dates for usable data were chosen to be 1/1/2006 and 12/27/2008 respectively. This gives us 3 complete years of data. A week begins with the first shift on Monday and is assigned a weekday number 1, thus making Sunday the last day (7). To ensure that there are no incomplete weeks in the year, the years were defined as shown in Table 1.

Table 1: Definition of weeks and years

Year	Start date	End date	No of weeks
2006	1/2/2006	12/31/2006	52
2007	1/1/2007	12/30/2007	52
2008	12/31/2007	12/28/2008	52

As per the hospital rules, the employee leaves can be sorted into as many as 15 categories. For the purpose of this study, the categories of leaves that were included were Call in (Sick), Call in (Emergency) and Leave early (Sick). These three counts of leaves were grouped together and cumulatively referred to as the Sick leave usage as they were identified to represent the category of leaves which were unexpected and create last minute staff scarcities. Due to the varying lengths of shifts, the ideal unit for leave measurement was found to be hours instead of the number of days or number of shifts.

3.2 Holiday analysis

The patterns for the weekly sick leave usage for the 3 years were studied and the trends were compared with each other.

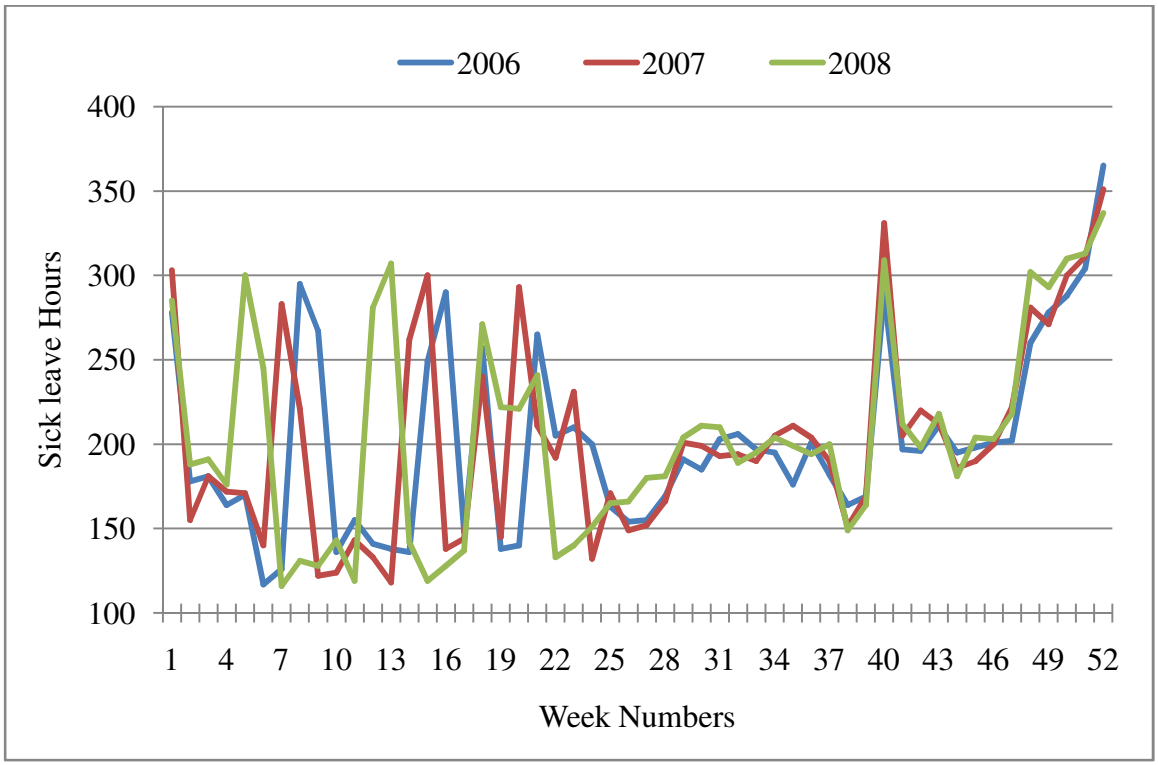

Figure 5: Weekly sick leave usage

Figure 5 shows the weekly sick leave usage for the 52 weeks in each of the 3 years 2006, 2007 and 2008. Ignoring the peaks, we note that the usage stays on the lower side after the first month of the year and then it begins to surge towards the beginning of the second half of the year and peaks in the last months. If we look carefully at the peaks in the first half of the year, we see that they occur during different parts of the year for different years. Thus, it seems prudent to trace the exact reasons for the rise and fall in sick leave usage over the year.

Table 2: Holiday calendar

Holiday Name	Occurrence	Week numbers affected		
		2006	2007	2008
New Year's Day	Jan 1	1	1	1
Epiphany	Jan 6	1	1	1
Carnival	7 weeks before Easter Monday	8 - 9	7 - 8	5 - 6
Good Friday	Week before Easter Monday	15	14	12
Easter Monday	Day after Easter Sunday	16	15	13
Labor Day	May 1	18	18	18
Ascension Day	39 days after Easter Sunday	21	20	18
Whit Sunday	50 days after Easter Sunday	22	21	19
Whit Monday	51 days after Easter Sunday	23	22	20
Corpus Christi	60 days after Easter Sunday	24	23	21
Unity Day	October 3	40	40	40
All Saints Day	November 1	44	44	44
Repentance day	Wednesday before Nov 23	47	47	47
Christmas Day	Dec 25	52	52	52

The analysis leads to the conclusion that there are specific holidays and festivals which account for the behavioral patterns of the weekly sick leave usage. These factors were identified and implemented in the design of the model. While some of these factors are public holidays and festivals, there are also some seasonal factors that have a deep impact on the sick leave usage patterns. The holiday calendar in Table 2 gives us an idea of their occurrences. The holidays can be categorized into fixed and moving holidays. A fixed holiday is one that occurs on the same day every year. These are usually the political holidays which are observed on specific days. The bulk of the holidays fall under the moving holidays section. A lot of the religious festivals like Easter and the ones related to it form the crux of the moving holidays of the year. It is important to verify that each of these factors affects the sick leave usage in a consistent manner over the years so that they can be used as an input in the forecasting model. This section analyzes the effects of each of these holidays and seasonal factors on the sick leave usage.

The sick leave usage in the first few weeks of the year are affected by the occurrence of the New Year and the Epiphany holidays. These holidays are fixed holidays. The sick leave usage is very high during the first week of the year. Even though the usage drops in the weeks following the New Year and Epiphany, it is still relatively high for the next 3-4 weeks after which it drops until the arrival of the Carnival holidays. Figure 6 shows the effect of the New Year and Epiphany holidays on the sick leave usage. The reason for the 'Epiphany + 4' week following a different trend in 2008

is because of the early advent of the Carnival festival which overlaps with that week in that year alone.

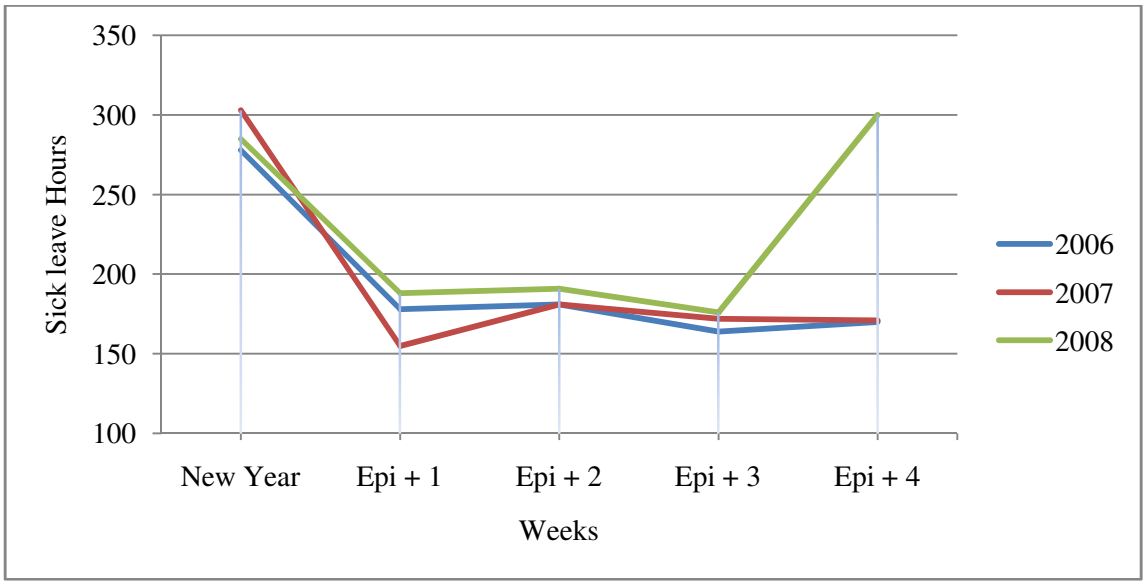

Figure 6: New Year and Epiphany effects

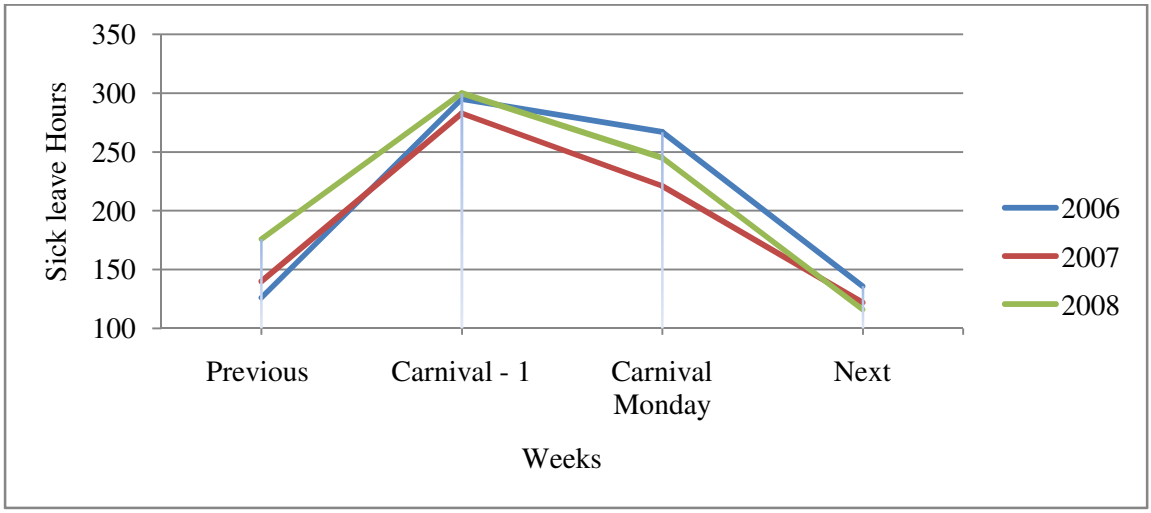

Figure 7: Carnival effects

Carnival Monday or Rose Monday occurs 7 weeks before Easter Monday and the festival is celebrated for a week starting from the previous Thursday through the next Wednesday. Due to the moving nature of Easter, this festival also falls under the moving holiday category. Sick leave usage rises during the 2 weeks of Carnival with the first week showing slightly higher numbers than the second week. Figure 7 shows the patterns observed during this time of the year.

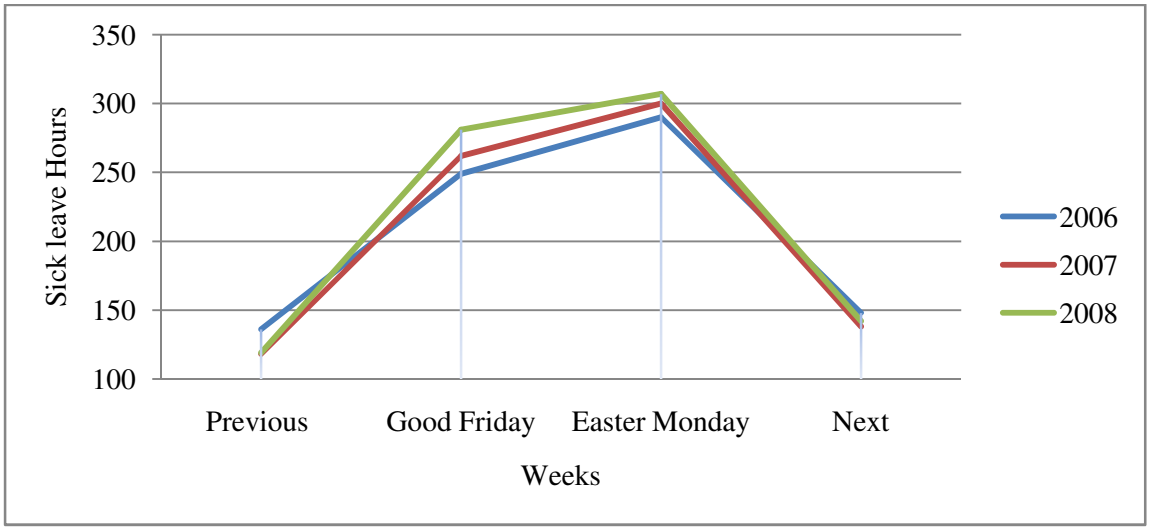

Figure 8: Easter effects

The two weeks affected by the Easter festival are the Good Friday week and the Easter Monday week, where the sick leave usage hours increase significantly. As mentioned earlier, Easter and Good Friday are moving holidays too. Figure 8 depicts the Easter time sick leave usage trends. The May Day or the Labor Day week also attracts a heavy usage of sick leaves as seen in the curves of Figure 9. This is a fixed holiday and always occurs during the same time of the year. The reason for the higher sick leave usage in the week following the May Day week in 2008 is because of the occurrence of

the Ascension Day/ Whit week effects during this week. These are observed during later weeks in case of 2006 and 2007.

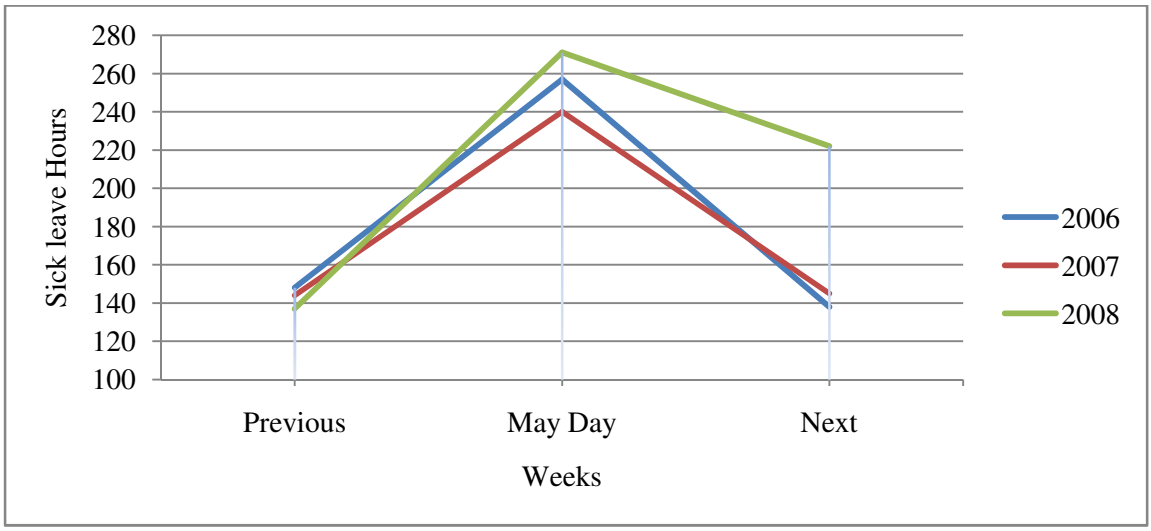

Figure 9: May Day effects

Ascension Day, Whit week and Corpus Christi are moving holidays whose occurrence remains constant relative to Easter Day. The sick leave usage shows a surge during the festival weeks and dips again after the Corpus Christi week. Figure 10 gives us an idea of the sick leave usage during this period. Unity Day and All Saints Day are fixed holidays and by coincidence, fall in weeks 40 and 44 respectively in all of the 3 years that are being considered for this analysis. While Unity Day results in a rise of sick leave usage, All Saints day shows a slight fall in the sick leave hours as evident from Figure 11.

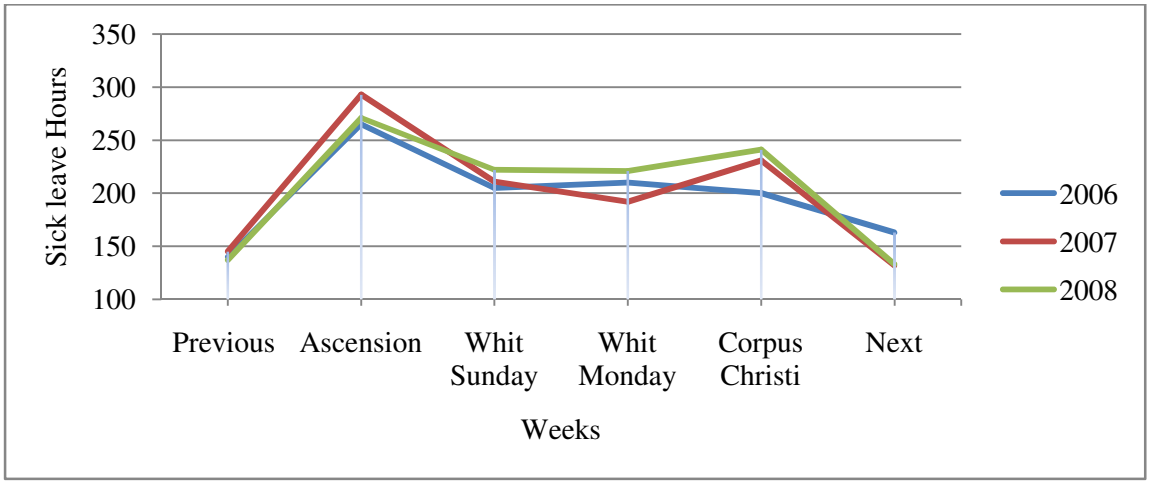

Figure 10: Ascension Day, Whit week and Corpus Christi effects

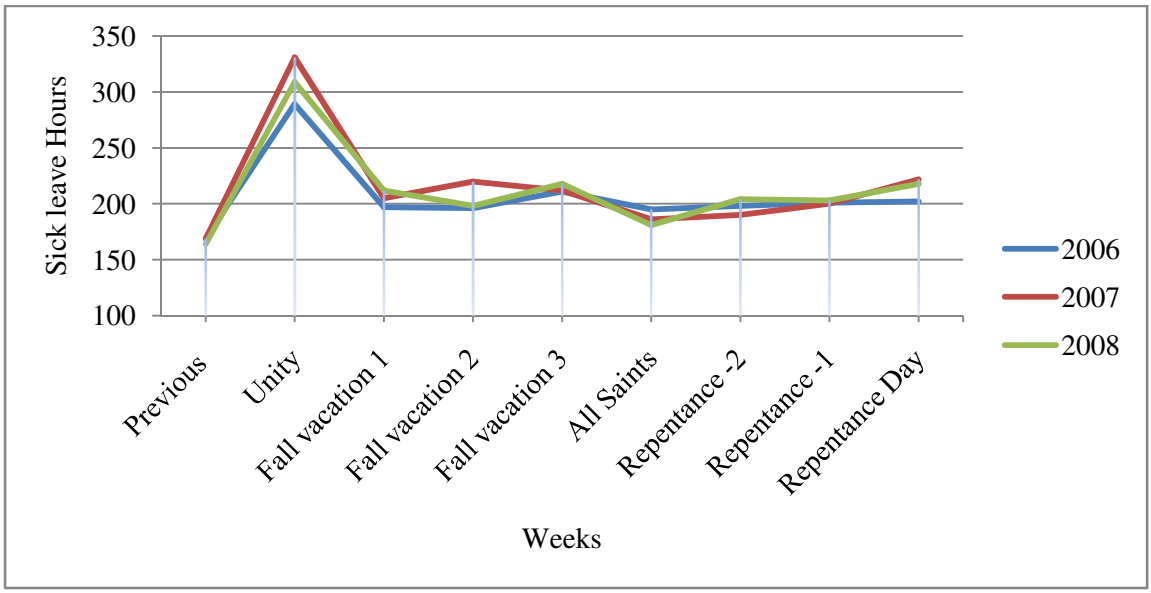

Figure 11: Unity, All Saints, Repentance and Fall vacations effects

Even though the Repentance Day does not fall on the same day every year, it is considered a fixed holiday as it roughly falls in the same time every year. The Repentance week and the 2 weeks leading to it generally show larger numbers due to the

seasonality effect experienced towards the last 2 months of the year. Figure 11 covers the sick leave usage details for the weeks around Repentance Day.

Sick leave usage sees a mighty surge towards the end of the year. This could be because the nurses are trying to use up the leaves left in their quota before the end of the year. Christmas vacations also have a major role to play in governing sick leave usage, with the usage rapidly rising in the 4 weeks leading to Christmas. Figure 12 gives a clear picture of how the approaching Christmas holidays affect leave usage. It was observed that the week containing the Christmas vacation experiences the highest sick leave usage for the year in all 3 years, the hourly count being nearly 75% larger than the yearly average.

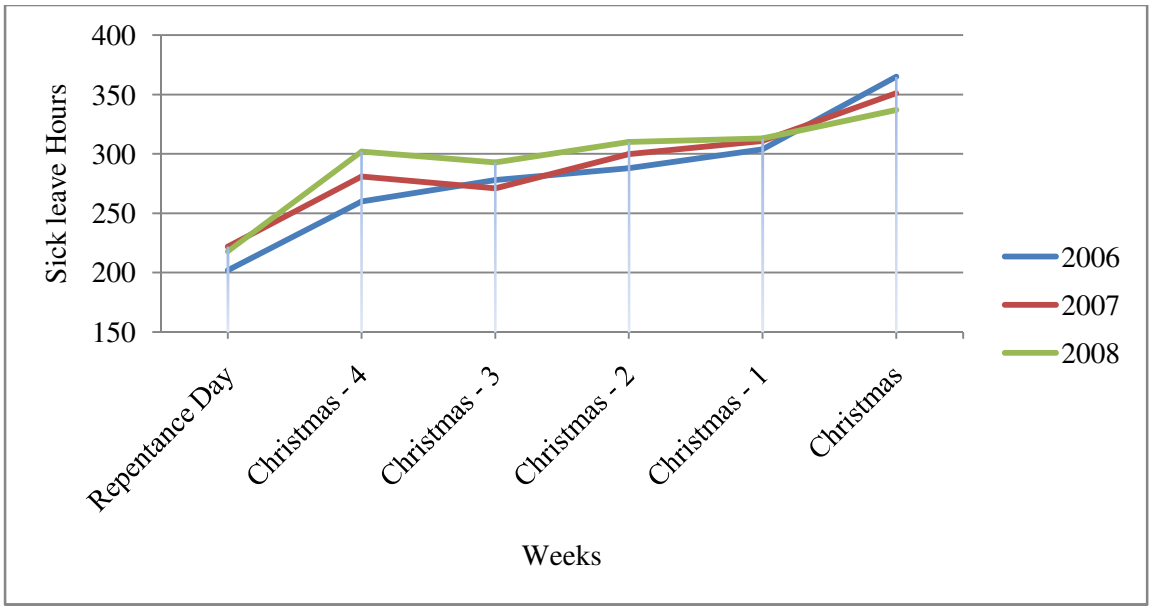

Figure 12: Christmas effects

The summer school vacation usually starts around late June and extends through the months of July and August. The patterns during this time period can be divided into 3 sections. The first phase is where the usage shows a slight increase when compared to the usages in the earlier weeks which were normal or unaffected by holidays. The middle phase is where the usage of leaves form the highest level of the plateau. The usage dips after the end of the summer vacations and stays low up to the advent of the fall vacations. The fall vacations' effects usually last for about 2 or 3 weeks in the month of October and November where they are squeezed between the occurrences of the Unity day and the All Saints Day holiday. The usage during this period is relatively higher than a normal week. Figure 13 and Figure 11 throw some light on the summer and fall vacation effects, respectively.

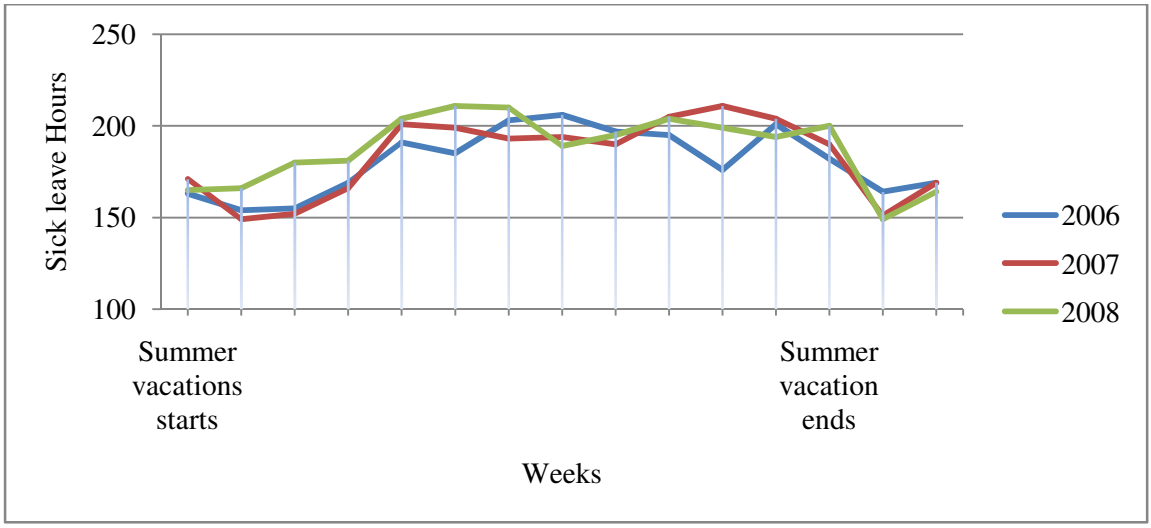

Figure 13: Summer vacations effects

4. Proposed model

4.1 ANN model construction

The construction of the model started with the allocation of parameter values. In our model, the factors identified to be contributing to the sick leave usage were the holidays, festivals and seasonality. Each of these factors demanded a separate node and hence, these form the 25 input nodes of the model. The output we desire is the number of hours lost to sick leaves during the week. Hence this forms the solitary output node. It is possible to include other outputs which are affected by the same set of input nodes, but our problem concerns just the weekly sick leave usage.

In our problem, every week is a data point with a factor value entered into each input node for that week and the week's sick leave usage (hours) forming the value of the output node. Based on the availability of data, our training data points ended up being the 104 weeks in the years 2006 and 2007, while the forecasting data points were the 52 weeks of the forecast year 2008. The cross validation cycle was bypassed because of the limited availability of data and hence, no data points were assigned for cross validation.

For arriving at the number of hidden layers and hidden nodes, we used a combination of suggestions from the literature review and trial and error experimentation. In our model, we used two hidden layers because of the large number of input nodes. Even though the accuracy does not improve greatly when moving from one hidden layer to two, this construction keeps the model the flexibility to handle more complicated data patterns. It was observed that as we increase the number of hidden nodes, the time taken for training process also increases significantly. In our problem, we found out that there

was no significant impact on the results by varying the number of hidden nodes and this is in agreement with what we found through our literature review. Our forecasting model has 4 nodes in each hidden layer, and this architecture's performance is equally satisfactory as any other reasonable choice for number of hidden nodes.

The performance of the model was tested for different runs with varying number of iterations with mean squared error as the performance measure. Every time the model is retrained, the arc weights take a different path to optimization and hence, might lead to differences in forecasted values for two different runs consisting of the same number of iterations. For a large number of iterations, the model's forecasted values across different runs are much more consistent as the transfer function gets closer to attaining the best arc weights. It was observed that the results did not improve greatly for a larger number of iterations and a low mean squared error was achieved rather quickly. The details of this parameter are discussed in the Results section.

4.2 Building the input matrix

The matrix formed by the 25 input columns and the 52 week rows is referred to as the input matrix. The values that are entered into the cells of this matrix are called components. The components are entered into the input nodes by following a step by step procedure. The default component for all cells is 0. If a particular holiday effect occurs during a particular week, we enter a 1 in the cell of that week against the column that represents that holiday effect.

W	NY	Ep	K1	K2	GF	E	M	A	W1	W2	CC	SF1	SF2	SF3	U	HF	AS	R2	R1	R	C4	C3	C2	C1	C
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 14: Input matrix for year 2007

Table 3: Expansion for column names used in input matrix

Column Numbers	Symbols	Holidays/ Festivals
1	NY	New Year's Day/Eve
2	Ep	Epiphany and the 4 weeks following it
3,4	K1, K2	2 weeks of Carnival
5,6	GF, E	Good Friday and easter holidays
7	M	May Day or Labor Day
8	A	Ascension
9,10	W1, W2	Whit week
11	CC	Corpus Christi
12,13,14	SF1, SF2, SF3	Summer vacations
15	U	Unity Day
16	HF	Fall vacations
17	AS	All Saints day
18, 19, 20	R2, R1, R	Repentance day and the 2 weeks before
21, 22, 23, 24, 25	C4, C3, C2, C1, C	Christmas and the 4 weeks before

We start filling out the matrix starting from the left most column. The order of the columns is not important and can be shuffled. In this model, the columns were arranged in chronological order of the occurrence of the event from left to right. Most of the entries are straightforward. For example, the first column is New Year which occurs in

the first week of the year. We proceed to mark a 1 in the cell that forms the intersection of the NY column and week 1 row.

The input matrix for the year 2007 is shown Figure 14. The expansion for all the column headings in Figure 14 has been provided in Table 3. The presence of a 1 in a particular column can be interpreted as a signal for the model to look for similar weeks in its training data set, i.e., weeks which have a 1 entered in the same column. The actual week number or year does not form an input node in this model even though these are popular inputs in most time series models. It was discovered that the week number and year are not useful inputs for this particular problem and might also end up aggravating the performance. To demonstrate this with an example, we look into the calendar which tells us that Easter falls in week 15 in 2007 and week 13 in 2008. Had the week number been an input node, the model would see a connection between week 15 and Easter week's sick leave usage during training and would try to use that information while forecasting the week 15 sick leave usage for 2008. Clearly, this spurious connection is detrimental to accurate forecasting.

4.3 ANN behavior

As we have seen, the component of each cell is either 0 or 1. To determine the effect that the value of the component has on the performance of the model, a simple test was performed. Consider a simple model with 1 input and 1 output node. The model was first trained in two separate cycles called Training cycle A and Training cycle B, the details of which are provided in Table 4.

Training cycle A has two data points while cycle B has three data points. Each training cycle was followed by a forecasting cycle. The model was used to forecast for 21 input data points with component values starting from 3.5 and extending up to 5.5 with an interval of 0.1 between each test point. Figure 15 shows that while the model tried to capture the trend between component values 4 and 5, any value outside this range produced an output very close to the output produced by the nearest boundary of the component value's range defined in the training cycle. This indicates that the ANN can forecast satisfactorily only within the range defined by the maximum and minimum component values used while training the model. For any component value outside this range, the forecasted output is going to be very close to the forecasted output for the nearest component value boundary.

Table 4: Data to study ANN behavior

	Training Cycle A	Training Cycle B
Input	Output	Output
4	100	100
4	-	120
5	200	200

If we look at the forecasted values for components 4 and 5 in the two cases, we observe that Case A produces results that directly compare with the training data. The

reason why the values forecasted are not exactly the same as the values the model was trained for is because the ANN introduces a little bit of variability into the forecasts and this can be viewed as beneficial for most forecasting applications.

In Case B, the forecasted value for a component of 5 is around 200, which corresponds to what it was trained for. The forecasted value for a component of 4 is around 110. This can be explained by the fact that the model was trained for the component value 4 twice, and so it ends up roughly averaging the two corresponding output values in the training data.

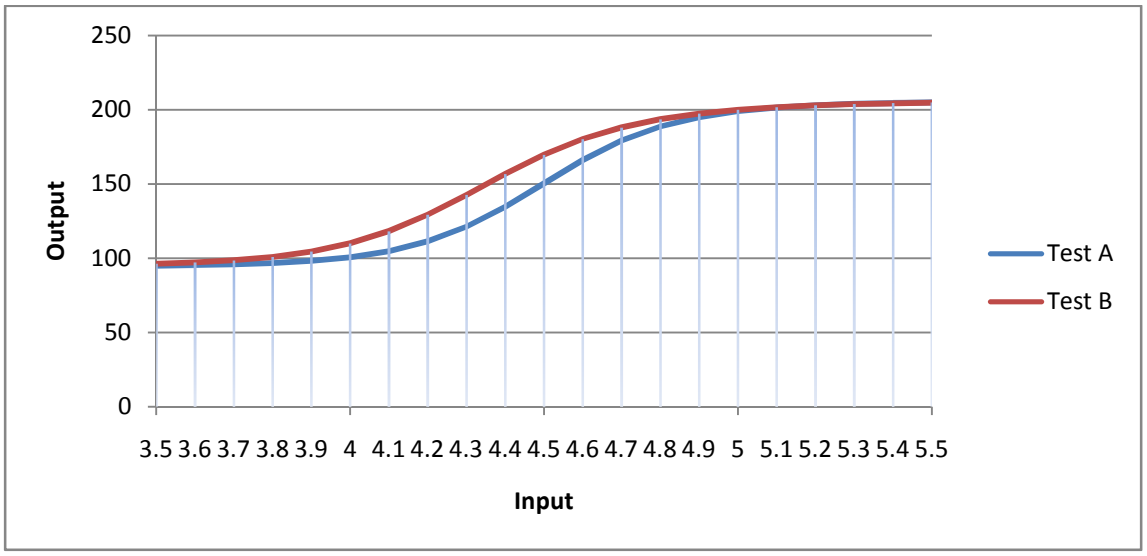

Figure 15: Forecast results for cases A and B

5. Results

The software used for this forecasting exercise was Neurosolutions version 5. The training was carried out for various numbers of iterations. It was observed that the MSE is reduced drastically during the first few iterations of the training cycle as it is evident from the steepness of the curve in Figure 16. Although Figure 16 depicts only one such training run, the basic trends are similar across all runs. The training curve starts stabilizing around the 1000th iteration and the improvement in the MSE is rather small after that. The MSE value at iteration No.3000 is 0.002656 which improves to a value less than 0.0024 for a higher number of iterations. It is always advisable to stop the optimization abruptly once the curve shows signs of stabilization. This is to avoid overfitting problems that can be caused by further reductions in the MSE. The model saves the final arc weights into its transfer function and we now have the transfer function which is used for forecasting.

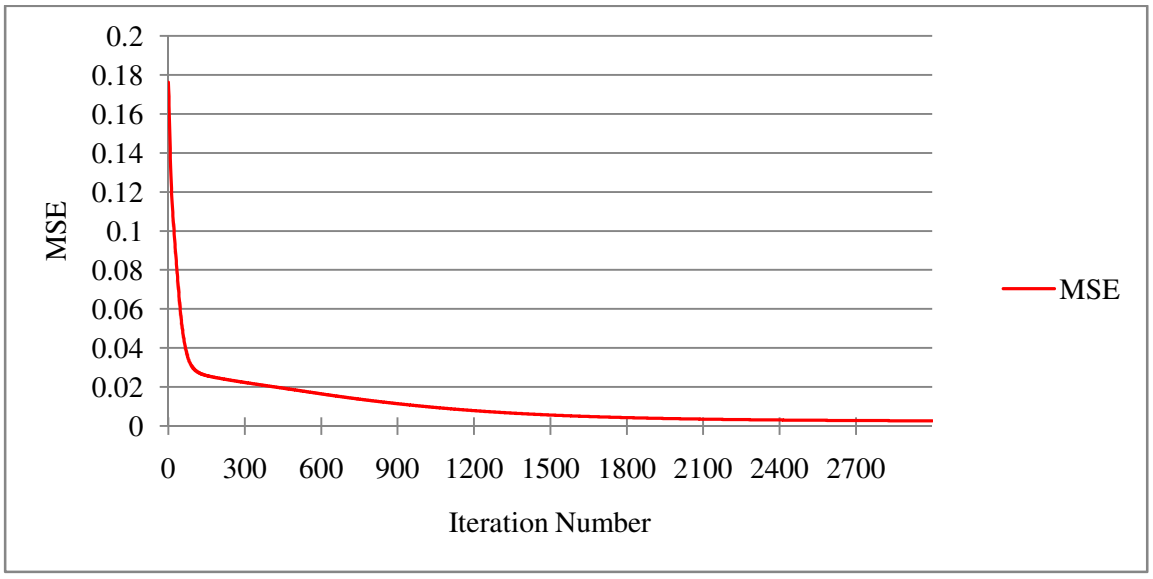

Figure 16: Training cycle progress

The forecasting cycle was run and the results obtained are shown in Figure 17. These are the results obtained after iteration number 10000. The mean squared error of this forecast is 175.3.

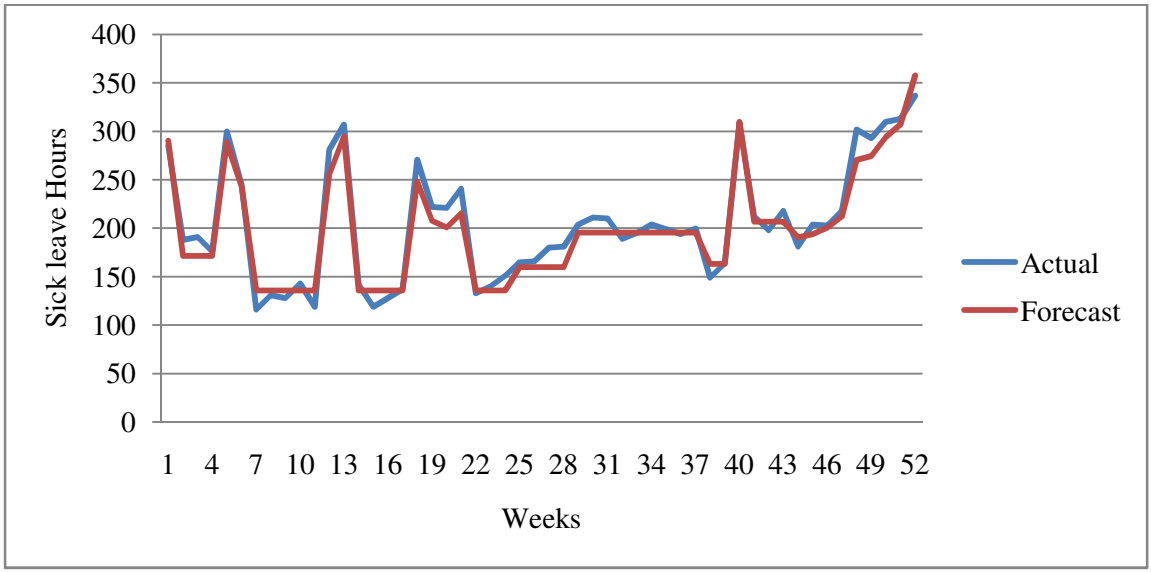

Figure 17: Actual and forecasted outputs for year 2008

Table 5 gives the statistics for various numbers of iteration in a run of 10000 iterations. A forecasting cycle is conducted at each of these checkpoints and the results are tabulated. We see that the improvement in the forecast performance is insignificant after we surpass a certain number of iterations. In our problem, it is safe to say that number is approximately 3000 iterations.

Table 5: Iteration checkpoints

Number of Iterations	Final MSE of training data	MSE of Forecast	Minimum absolute Error	Maximum absolute error
500	0.0152	916.12	0.035	62.5
1000	0.0084	558.60	0.156	49.9
1500	0.0049	358.14	0.130	39.7
2000	0.0034	265.43	0.377	34.6
3000	0.0026	202.28	0.325	32.7
5000	0.0024	179.61	0.156	31.7
10000	0.0024	175.30	0.721	31.5

Equally important is the precision in forecasting, and an experiment was conducted to determine the number of iterations at which the forecast becomes stable. The results are displayed in Table 6. In this experiment, the training and forecasting MSE were compared at various iteration checkpoints for 3 different runs of the training cycle. The forecast MSE suggests that the model's results are close to consistent when it is run for up to 5000 iterations but a 10000 iteration run produces an extremely high level of precision in forecasting.

Table 6: Experiment for testing consistency of forecast results

Number of iterations	Run Number 1		Run Number 2		Run Number 3	
	Training MSE	Forecast MSE	Training MSE	Forecast MSE	Training MSE	Forecast MSE
2000	0.0031	243.73	0.0062	439.65	0.0075	403.11
3000	0.0025	191.60	0.0026	207.78	0.0027	216.26
5000	0.0024	181.63	0.0024	183.23	0.0024	183.55
10000	0.0024	175.18	0.0024	175.18	0.0024	175.36

6. Conclusion

An ANN based forecasting model was thus successfully built to predict the weekly usage of sick leaves over the year. The various parameters chosen for the construction of the model seemed to perform satisfactorily. This model can be used to fine tune the staffing and scheduling models for the hospital. By having some insight into the number of unexpected or last minute absences, the nurse managers can decide on the overstaffing patterns or the number or on-call nurses to appoint for the week. This way, they will be better prepared to deal with the crisis in the most economical way possible.

The results produced by ANN have not been compared with any traditional methods of forecasting in this report and this creates a lot of scope for future research. The same data can be applied to models built on other forecasting techniques and the performance of ANN can be compared to these other techniques.

7. References

T. Al-Saba and I. M. El-Amin. Artificial neural networks as applied to long-term demand forecasting. *Journal of Artificial Intelligence in Engineering*, 13(2):189-197, 1999.

J. Carvel. Nurses top public sector sick leave table. *guardian.co.uk,* June 2005.

K. Chakraborty, K. Mehrotra, C. K. Mohan and S. Ranka. Forecasting the behavior of multivariate time series using neural networks. *Neural Networks,* 5:961–970, 1992.

H. Chen, C.A. Canizares and A. Singh. ANN-based short-term load forecasting in electricity markets. *Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference*, 2:411–415, 2001.

R. D. Hackett, P. Bycio and R. M. Guion. Absenteeism among hospital nurses: An Idiographic-Longitudinal Analysis. *The Academy of Management Journal*, 32(2):424-453, 1989.

K. Hornik, M. Stinchcombe and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks,* 2:359–366, 1989.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4:251–257, 1991.

K. Hornik. Some new results on neural network approximation. *Neural Networks* 6:1069–1072, 1993.

B. Irie and S. Miyake. Capabilities of three-layered perceptrons. In: *Proceedings of the IEEE International Conference on Neural Networks*, I, pp. 641–648, 1988.

S. Kang. An investigation of the use of feedforward neural networks for forecasting. Ph.D. Thesis, Kent State University, 1991.

N. Kohzadi, M. S. Boyd, B. Kermanshahi and I. Kaastra. A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing*, 10:169–181, 1996.

G. Lachtermacher and J. D. Fuller. Back propagation in time series forecasting. *Journal of Forecasting,* 14:381–393, 1995.

R. P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, April:4–22, 1987.

D. Srinivasan, A. C. Liew and C. S. Chang. A neural network short-term load forecaster. *Electric Power Systems Research,* 28:227–234, 1994.

Z. Tang, C. Almeida and P. A. Fishwick. Time series forecasting using neural networks vs Box-Jenkins methodology. *Simulation* 57(5):303–310, 1991.

Z. Tang and P. A. Fishwick. Feedforward neural nets as models for time series forecasting. *ORSA Journal on Computing,* 5(4):374–385, 1993

G. Zhang, B.E. Patuwo, Y.M. Hu. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting,* 14:35-62, 1998.

X. Zhang. Time series analysis and prediction by neural networks. *Optimization Methods and Software,* 4:151–170, 1994.

Artificial neural network. (n.d.). In *Wikipedia.* Retrieved October 19, 2010, from http://en.wikipedia.org/wiki/Artificial_neural_network

Vita

Srikanth Tondukulam Seetharaman was born in Bangalore, India. After completing his work at National Public School, Bangalore in 2002, he entered the National Institute of Technology, Tiruchirappalli and received the degree of Bachelor of Technology in Production Engineering in May 2006. During the following years, he was employed as an R&D Engineer with Larsen & Toubro Limited. In August 2008, he entered the Graduate School at the University of Texas at Austin to pursue his Masters degree in Operations Research and Industrial Engineering.

Email: sriknitt@gmail.com

This report was typed by the author.