

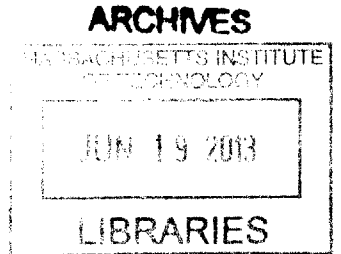
Absenteeism Prediction and Labor Force Optimization in Rail Dispatcher Scheduling

By

Taylor Jensen
B.S. Construction Management, Brigham Young University, 2008

and

Qi Sun
B.S. Computer Science and Technology, Qingdao University, 2004



Submitted to the Engineering Systems Division in Partial Fulfillment of the
Requirements for the Degree of

Master of Engineering in Logistics

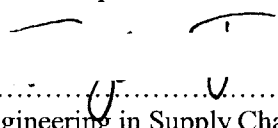
at the

Massachusetts Institute of Technology


June 2013

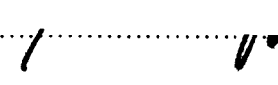
© 2013 Taylor Jensen and Qi Sun
All rights reserved.

The authors hereby grant to MIT permission to reproduce and distribute publicly paper and electronic
copies of this document in whole or in part.

Signature of Author..........
Master of Engineering in Supply Chain Management Program, Engineering Systems Division
May 10, 2013

Signature of Author.....
Master of Engineering in Supply Chain Management Program, Engineering Systems Division
May 10, 2013

Certified by..........
Dr. Anthony J. Craig
Postdoctoral Associate
Thesis Supervisor

Accepted by..........
Prof. Yossi Sheffi
Elisha Gray II Professor of Engineering Systems, MIT
Director, MIT Center for Transportation and Logistics
Professor, Civil and Environmental Engineering, MIT

Absenteeism Prediction and Labor Force Optimization in Rail Dispatcher Scheduling

by
Taylor Jensen and Qi Sun

Submitted to the Engineering Systems Division in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering in Logistics

Abstract

Unplanned employee absences are estimated to account for a loss of 3% of scheduled labor hours. This can be costly in railroad dispatcher scheduling because every absence must be filled through overtime or a qualified extra dispatcher. One factor that complicates this problem is the uncertainty of unplanned employee absences. The ability to predict unplanned absences would facilitate effective scheduling of extra dispatchers and help reduce overtime costs. This thesis uses data from a railroad company over a four year period to examine company-wide factors thought to impact the number of unplanned absences among dispatchers. Using Poisson regression, we identify several factors that provide statistical evidence of influencing the number of unplanned absences. These factors are month, snowstorms, shift, and certain holidays. Despite these findings, the overall predictive capability of our regression model is very weak. Instead, we model the number of unplanned absences by shift as a random process with a Negative Binomial distribution and use Monte Carlo simulation to explore the impact on overtime costs of increasing the number of scheduled extra dispatchers and increasing the number of positions on which each employee is qualified to work. Our results show that increasing the number of extra dispatchers has a greater effect on reducing overtime, but the cost savings from reducing overtime expenses are not enough to offset the additional labor costs of having more employees on staff. Our results provide insight regarding the relationship among extra staff, higher levels of qualification among employees, and the willingness to use overtime in handling unplanned absences.

Thesis supervisor: Anthony J. Craig
Title: Postdoctoral Associate

Table of Contents

List of Tables.....	5
List of Figures	6
1 Dispatcher Scheduling in the Railroad Industry.....	7
1.1 Scheduling at RailCo.....	7
1.2 Motivation.....	8
2 Literature Review	10
2.1 Causal Factors of Employee Absenteeism.....	10
2.2 Scheduling Replacements: The Assignment Problem.....	12
3 Methods.....	13
3.1 Evaluating Count Data: Introduction	15
3.1.1 The Binomial Distribution.....	15
3.1.2 The Poisson Distribution	16
3.1.3 The Negative Binomial Distribution.....	17
3.1.4 Goodness of Fit Tests	18
3.2 Model Selection.....	19
3.3 Generalized Linear Regression Models.....	21
3.3.1 Goodness of Fit Tests for Generalized Linear Regression Models	22
3.4 Inputs for Simulation.....	22
3.4.1 Distributions of Employee Qualifications	23
3.5 Optimization	28
3.6 Simulation	32
4 Data Analysis: Regression.....	35
4.1 Statistically Significant Parameters.....	37
4.1.1 Month.....	37
4.1.2 Shift.....	38
4.1.3 Holidays.....	38
4.1.4 Snowstorms.....	40

4.1.5	Planned Absences	41
4.2	Statistically Insignificant Factors	41
4.2.1	Hunting Season.....	41
4.2.2	Football Games.....	42
4.2.3	Day of the week and Day of the Month	42
4.3	Marginal Effects	44
4.4	Significance of Model.....	45
4.5	Goodness of Fit for Negative Binomial Regression	46
4.6	Other Considerations: Yearly Trends and Absences by Employee.....	47
5	Data Analysis: Simulation.....	48
5.1	Slide Cost.....	50
5.2	Overtime Cost.....	51
5.3	Extra Cost	53
5.4	Current Cost.....	55
6	Conclusion	56
6.1	Other Considerations in Employee Staffing	57
6.2	Future Research	60
7	Bibliography	61

List of Tables

Table 1: Shift start and end times	13
Table 2: Sample absence codes	14
Table 3: Goodness of fit tests for absences by shift	20
Table 4: Statistics for distributions of extra board qualifications	26
Table 5: Statistics for distributions of incumbent qualifications	28
Table 6: Number of positions for each day and shift	34
Table 7: Mean and standard deviation of each employee type	34
Table 8: Average qualifications per employee	35
Table 9: Possible factors that influence unplanned absences	36
Table 10: Regression statistics for month	37
Table 11: Regression statistics for shift	38
Table 12: Nine most common paid Holidays	39
Table 13: Regression statistics for Holidays	40
Table 14: Regression statistics for snowstorms	40
Table 15: Regression statistics for planned absences	41
Table 16: Regression statistics for hunting season	42
Table 17: Regression statistics for football games	42
Table 18: Regression statistics for days of the month	43
Table 19: Regression statistics for day of the week	44
Table 20: Summary of actual effects of statistically significant parameters	45
Table 21: Comparison of results from Poisson and Negative Binomial regression	46
Table 22: Average planned absences	49

List of Figures

Figure 1: Example positions by day and shift.....	7
Figure 2: Irregularity of unplanned absences in 2012	9
Figure 3: Probability Distribution of Absences	15
Figure 4: Sample Binomial Distributions based on varying values of n and p	16
Figure 5: Sample Poisson distributions based on varying values of λ	17
Figure 6: Sample Negative Binomial distributions based on varying values of r and p	18
Figure 7: Distribution of unplanned absences compared to sample distributions	19
Figure 8: Samples from data tables	24
Figure 9: Distributions of qualifications for extra board employees.....	25
Figure 10: Distributions of qualifications for incumbent employees.....	27
Figure 11: Qualification Matrix.....	29
Figure 12: Cost Matrix	30
Figure 13: Solution Matrix.....	32
Figure 14: Total absences by year	47
Figure 15: Cumulative absences by year.....	48
Figure 16: Slide cost.....	50
Figure 17: Overtime cost	51
Figure 18: Overtime cost by qualifications	53
Figure 19: Extra cost.....	54
Figure 20: Slide cost, overtime cost, and extra cost	55
Figure 21: Effect of changing extra board size or qualification.....	56
Figure 22: Total labor cost	58
Figure 23: Total labor cost, extra cost, and extra board cost.....	59

1 Dispatcher Scheduling in the Railroad Industry

RailCo¹ operates several thousand miles of track across the United States. The department that directs traffic across this network employs over four hundred dispatchers and operates 24 hours a day, 7 days a week, 365 days a year. The daily assignment of these four hundred dispatchers across their unique positions combined with the scheduling of employee vacations requires the labor of eight full-time employees. These scheduling employees are required to follow strict scheduling rules that govern how employees are qualified to work in specific positions, when employees can take vacation time, and how employees are disciplined for being absent from work.

1.1 Scheduling at RailCo

RailCo has three shifts of approximately ninety positions that must be staffed every day. Each of these positions is associated with a length of track over which the dispatcher directs railroad traffic. Before a dispatcher can work on a position he must be trained and receive the corresponding qualification for that position. The dispatcher that is regularly assigned to a certain position is called the “incumbent” for that position. Figure 1 shows an example of how positions are allocated by day and shift.

	Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
1st Shift (6:30-14:30)	Position 1	Position 1	Position 1	Position 1	Position 1	Position 4	Position 4
2nd Shift (14:30-22:30)	Position 4	Position 4	Position 2	Position 2	Position 2	Position 2	Position 2
3rd Shift (22:30-6:30)	Position 3	Position 3	Position 4	Position 5	Position 3	Position 3	Position 3

Figure 1: Example positions by day and shift

¹ Alias, name has been changed

If an incumbent for any position is absent from work, for either planned or unplanned reasons, his/her position must be staffed by an alternate employee who has been pre-qualified to work in that position. RailCo maintains a group of extra employees without regular assignments, called “extra board,” that can fill in these incumbent vacancies.

The RailCo dispatcher workforce is unionized and has strict rules that govern their schedules and work positions. If no extra board employee is available and qualified to staff an incumbent vacancy, then an incumbent from another position can be moved from his position in what is called a “slide.” The vacancy created by sliding the incumbent can then be filled by an extra board employee qualified on that position. This process of sliding employees is repeated until all positions are staffed with qualified employees. If it is not possible or feasible to fill vacancies by sliding employees, then RailCo can call an employee from home and pay him/her overtime to fill the position.

Because of union rules, each time an incumbent is moved from his/her regular position in a slide he/she must be paid time and one half for that day, and each time an employee is called from home he/she must be paid a full day and one half extra. Because of these union agreements, RailCo has little flexibility in how they make assignments and schedule their employees without incurring extra cost. This scheduling problem is further complicated by unplanned employee absences.

1.2 Motivation

The number of unplanned absences on any given day occurs in unpredictable patterns. On seemingly random days during the year there are an unusually high number of employees that call in sick. Figure 2 shows the irregularity of unplanned absences by day during 2012.

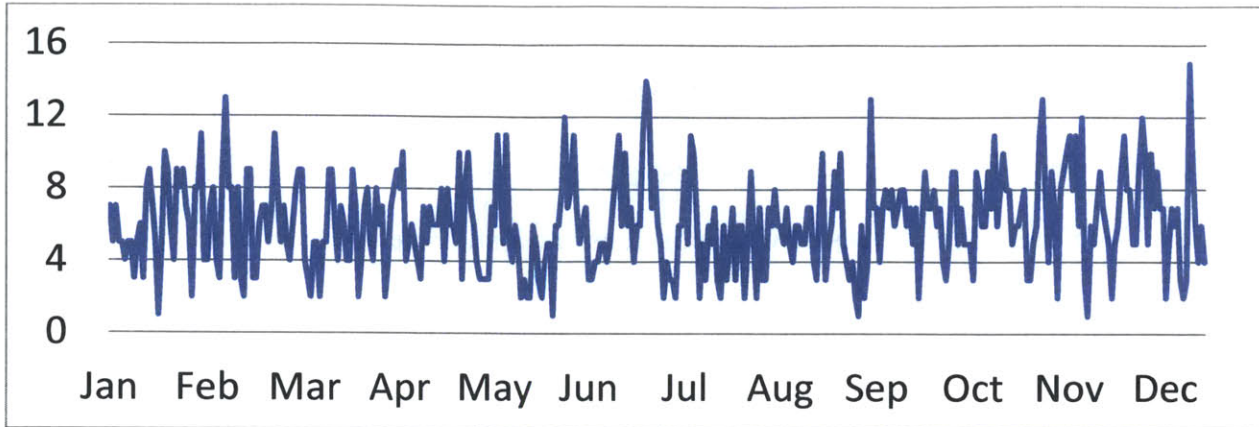


Figure 2: Irregularity of unplanned absences in 2012

Because RailCo cannot predict when these spikes in absenteeism will occur they are forced to keep a high number of extra board dispatchers on their payroll to cover worst case scenarios. Extra board dispatchers earn a full-time salary even if they do not have a specific assignment every day.

If RailCo could predict when days of unusually high absences would occur, they would be able to make adjustments to planned vacation allotments and training schedules in order to minimize employee slides and overtime pay. More accurate prediction of employee absences would also allow management to respond more quickly and accurately to employee requests for days off.

The first part of our research aims to answer the question of what factors influence unplanned dispatcher absences, and by using these factors is it possible to build a model that will predict how many absences will occur on any given day and shift. This knowledge will give RailCo the ability to make better decisions about staffing levels and make allotments for planned vacation days, and will have a significant impact on the both the profits and employee relations of RailCo.

A further part of our research will apply simulation techniques to help RailCo find the most appropriate number of extra board employees and the optimal number of qualifications that each

dispatcher on the extra board should have. Both of these factors will affect the total labor costs for RailCo. The more qualifications that each extra board dispatcher has, the more flexibility RailCo will have in their scheduling, and the less likely RailCo will be to incur the extra costs associated with sliding employees between positions or calling in people from home.

2 Literature Review

Most of the research conducted on employee absenteeism focuses on predicting absences based on the traits of the individual employee, or based on large scale economic factors that affect the entire population. Our research is somewhat unique in that it focuses on factors that affect absences on a company-wide scale. This includes factors such as shift, day of the week, month, holidays, as well as events that are specific to RailCo's geographical region. As RailCo gains the ability to predict absenteeism based on these factors, they will be able to customize the size of their labor force. As a background to our research, we will now review the literature that has been done on predicting absenteeism and the methods available for scheduling replacements for absent employees.

2.1 Causal Factors of Employee Absenteeism

In 2010 it was estimated that the total cost of absenteeism in the United States was \$118 billion (Weaver, 2010). A Mercer study estimated that the total costs of these unplanned absences were as high as 8.7% of payroll (Carpenter & Wyman, 2010). It has also been estimated that unplanned absences account for a loss of approximately 3% of scheduled labor hours (Bureau of Labor Statistics, 2011). The high cost of absenteeism has motivated studies aimed at predicting employee absenteeism in the healthcare, manufacturing, and service industries.

In 1998, a twenty year review of studies on absenteeism concluded that in the mid-term and long term, factors such as gender, age, health, and job satisfaction are all significant predictors of absenteeism (Harrison & Martocchio, 1998). Data published by the Bureau of Labor Statistics in support many of these conclusions; for example, in 2011, women were almost twice as likely to be absent from work as men, and older employees were found to have slightly higher instances of absenteeism than younger employees (Bureau of Labor Statistics, 2011).

Other studies have examined more limited predictive factors on absenteeism. One study (Hausknecht et. al, 2008) found a negative correlation between local unemployment rates and absenteeism. Another study, of 514 security guards, found that employees that perceive their employer as being unfair have slightly higher rates of absenteeism (De Boer et. al, 2002). Other studies show that unplanned absences are higher among union employees than non-union employees (Carpenter & Wyman, 2010; Chaudhury & Ng, 1992).

Several factors cause absences in the short term, the most common being illness. However, as many as half of unplanned absences can be attributed to factors other than illness, such as doctor appointments, problematic relationships, and vehicle repairs (Prater & Smith, 2011). Weather is another cause of absenteeism, although this is generally limited to regions where snow and other winter conditions are common (Bureau of Labor Statistics, 2012)

In our research we evaluated some of the factors detailed above, but most of our analysis focused on macro factors specific to RailCo's work conditions and geography. This is because RailCo's primary objective was to understand the causes of variation in unplanned absences for the company as a whole and not for individual employees.

2.2 Scheduling Replacements: The Assignment Problem

Whenever an incumbent for a position is absent for unplanned reasons, another employee with the appropriate qualification must be assigned to fill that vacancy. A substantial amount of literature exists that addresses how to solve this so-called assignment problem.

Kuhn (1955) defined the problem this way: "...personnel-assignment asks for the best assignment of a set of persons to a set of jobs, where the possible assignments are ranked by the total scores or ratings of the workers in the jobs to which they are assigned." Kuhn developed what he called the Hungarian Method, based on work done by D. König (1936) and E. Egerváry (1931), as a way to solve the assignment problem. In this method, employees and jobs are arranged in a matrix where the rows of the matrix represent employees, the columns represent jobs, and each row/column combination contains the corresponding value of that employee being assigned to that job. An algorithm is then applied to the matrix which produces the optimal solution based on the value in each cell of the matrix. This optimization method has been used by many researchers. For example James Munkres (1957) used it to solve transportation problems, and more recently, Hultberg and Cardoso (1997) employed this method to assign teachers to subjects in school

In our application of the assignment problem to RailCo's scheduling problem, we will create a matrix where the rows represent employees, the columns represent positions, and each cell describes the cost of that employee working in that position. Solving the assignment problem will then produce the minimum cost solution of assigning incumbents, extra board employees, and overtime workers to the positions.

3 Methods

In our analysis of employee absences we used data from January 1, 2009, to December 31, 2012; each absence during this four year period had a corresponding date, time, employee number and reason for the absence. As a precursor to analyzing these absences, we first had to determine the shift on which they occurred and if they were planned or unplanned.

Because dispatchers start work at different times during the day, shifts were divided into eight hour blocks based on the earliest start time possible under union rules, which is 5:00 AM. Table 1 shows the start and end times for each shift. All start times between 12:00 AM and 4:59 AM were assigned to the third shift of the previous day.

Table 1: Shift start and end times

	Start Time	End Time
Shift 1	5:00 AM	12:59 PM
Shift 2	1:00 PM	8:59 PM
Shift 3	9:00 PM	*4:59 AM

*the following day

The next step in our analysis was to determine how RailCo differentiated between planned and unplanned absences. RailCo uses several description codes in their scheduling operations to label each individual absence. A partial list of codes is shown in Table 2.

Table 2: Sample absence codes

Code	Description	Planned
AD	ALT DISCIPLINE	Yes
AO	ABSENT WITHOUT LEAVE	No
BL	BEREAVEMENT LEAVE	No
CB	COMPANY BUSINESS	Yes
FB	FMLA BIRTH	Yes
FD	FIT FOR DUTY--CAN'T WORK	No
FE	FAMILY EMERGENCY	No
FF	FMLA FAMILY	No
FH	EXEMPT DISP FLOAT HOLIDAYS	Yes
FI	FRML INVESTIGATION	No

Generally, an absence was considered “unplanned” if the scheduler was not aware the employee would not arrive more than 24 hours in advance and as a result would not have the opportunity to adjust schedules before the work day started. For series of unplanned absences from one employee lasting longer than one day, absences were counted as “unplanned” for the first five consecutive days, and then “planned” for any remaining days. This is because after several days of consecutive unplanned absences the schedulers would presumably be able to adjust their plans appropriately to account for the previously “unplanned” absence. By categorizing all absences by day and shift, we created 4 years x 365 days x 3 shifts = 4,383 day/shift combinations. A probability distribution of these absences is shown in Figure 3.

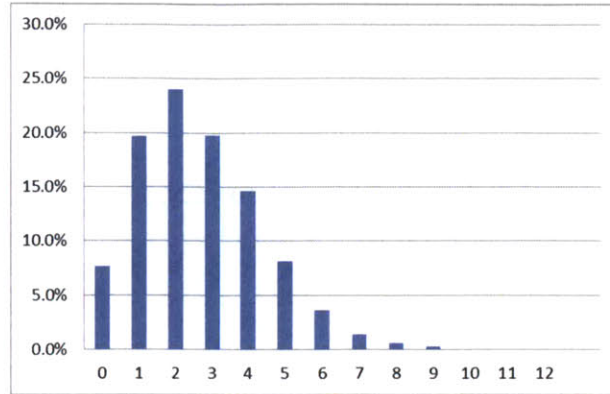


Figure 3: Probability Distribution of Absences

3.1 Evaluating Count Data: Introduction

The number of absences that occur on any given shift is categorized as count data, which are defined as values that are both non-negative and integers. There are several models that can be employed to evaluate count data; we will discuss the three most common, namely, Binomial, Poisson, and Negative Binomial.

3.1.1 The Binomial Distribution

The Binomial distribution is the simplest model used to evaluate count data arising from a series of independent and identical trials that result in either a success or a failure. The probability mass function for Binomial distributions is given by the equation:

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k} \quad (1)$$

Where n = the number of trials, p = the probability of success of each trial, and k = the number of successes. As an example of how this would be applied to RailCo, n would represent the number of employees that are scheduled to work on a given day or shift, p would represent the

probability that each employee would be absent from work, and k would represent the number of employees that were absent on that given day.

Figure 4 shows sample Binomial distributions based on varying values for n and p .

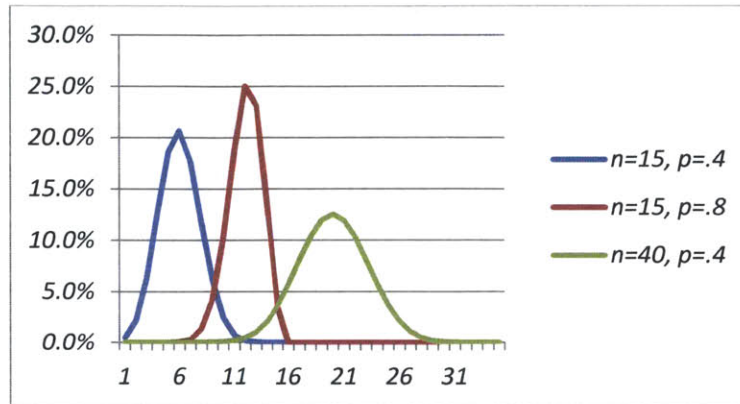


Figure 4: Sample Binomial Distributions based on varying values of n and p

This model is generally used for simple distributions where the probability of success is known. Because in our research the probability of success is unknown and varies from employee to employee, this model is not sufficient to analyze the absences that occur at RailCo.

3.1.2 The Poisson Distribution

The Poisson distribution is the most common model used to evaluate count data (Winkelmann, 2008). It is derived from the Binomial distribution as the number of trials increases towards infinity while holding the probability of success constant. The probability mass function of the Poisson distribution is:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (2)$$

Where $\lambda = E(X) = \text{Var}(X)$ and $e =$ the base of the natural logarithm (2.71828...). Example probability distributions based on different values of λ are shown Figure 5.

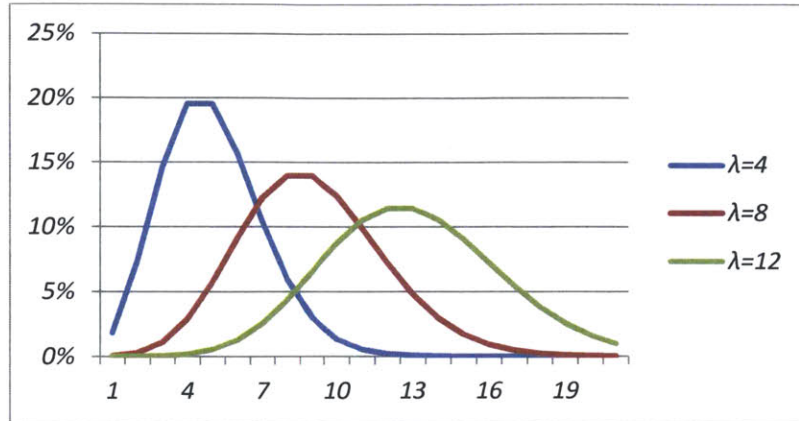


Figure 5: Sample Poisson distributions based on varying values of λ

The Poisson distribution is most useful if the mean and variance for the data being analyzed are the same. If the variance of a given set of data is greater than the mean, called over-dispersion, or smaller than the mean, called under-dispersion, then other models that are more flexible should be used. The amount of under/over-dispersion is simply the ratio between the variance and the mean. The distribution of unplanned absences by shift is slightly over-dispersed, so the Poisson distribution is not the best model for our data, but it will prove useful later in our analysis.

3.1.3 The Negative Binomial Distribution

The Negative Binomial distribution is a more flexible model than the Poisson distribution and is a good alternative for modeling count data that is over-dispersed (Winkelmann, 2008) because it allows the variance to take on a value different from the mean. The probability mass function for the Negative Binomial distribution is given by the equation:

$$P(X = k) = \frac{(k+r-1)!}{k!(r-1)!} (1-p)^r p^k \quad (3)$$

Where k = the number of successes, r = the number of failures, and p = the probability of a success.

The greater flexibility of the Negative Binomial distribution is illustrated in Figure 6, below, which shows three different distributions based on varying values of r and p .

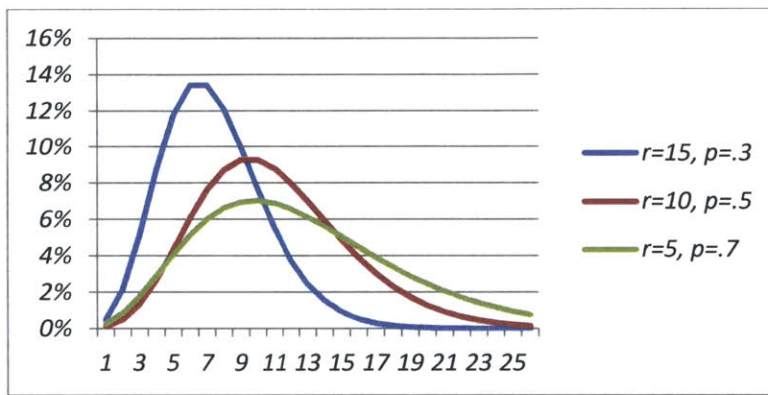


Figure 6: Sample Negative Binomial distributions based on varying values of r and p

We will use the Negative Binomial distribution extensively to model the data we evaluated in our research. The number of unplanned absences by shift, the number of extra board qualifications by day/shift, and the number of incumbent qualifications by day/shift can all be effectively modeled using Negative Binomial distributions.

3.1.4 Goodness of Fit Tests

Goodness of Fit Tests are used to determine how well a distribution approximates the sample data in question. A common goodness of fit test for the Poisson and Negative Binomial distributions is the Pearson Chi Square test, which is given by the equation:

$$\chi_{k-p-1}^2 = \sum_k \frac{(f_o - f_e)^2}{f_e} \quad (4)$$

Where f_o = the observed frequency, f_e = the expected frequency, k = the number of categories, and p = the number of parameters estimated from the data. We will use the Pearson Chi Square test to evaluate the goodness of fit for our data to both the Poisson and Negative Binomial distributions.

3.2 Model Selection

Using the mean and variance as parameters, we fit the distribution of absences to the Poisson distribution and the Negative Binomial distribution, and tested the degree of fit using the Pearson Chi Square test. A visual comparison of the distribution of absences by shift to the Poisson and Negative Binomial distributions is shown in Figure 7.

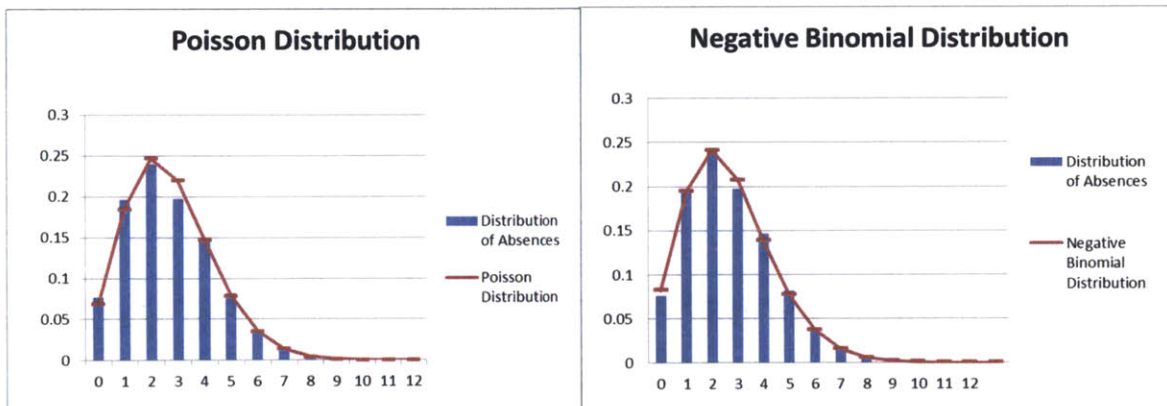


Figure 7: Distribution of unplanned absences compared to sample distributions

The Pearson Chi Square goodness of fit tests for our sample data are shown in Table 3.

Table 3: Goodness of fit tests for absences by shift

Negative Binomial

	Parameter	Estimate	Lower 95%	Upper 95%
Average	λ	2.6735	2.6228	2.7249
Overdispersion	σ	1.1110	1.0653	1.1599
	Chi Square	Prob<Chi Square		
Chi Square	9.3396	0.7468		

Poisson

	Parameter	Estimate	Lower 95%	Upper 95%
Average	λ	2.6735	2.6254	2.7222
	Chi Square	Prob<Chi Square		
Chi Square	43.7698	<.0001		

As seen in Table 3, the chi square value for the Poisson distribution is 43.77. Given that there are eleven degrees of freedom in this Chi Square test, the resulting probability value is less than .0001, from which we conclude the data is not from the Poisson distribution. The Chi Square value for the Negative Binomial distribution is 9.34, producing a probability value of .74, from which we conclude the Negative Binomial distribution is appropriate to model the number of unplanned absences.

Fitting the distribution of unplanned absences to distributions of known parameters such as Poisson and Negative Binomial allows us to use what are known as Generalized Linear Models to perform regression analysis. Poisson regression and Negative Binomial regression are both GLMs. Because Poisson regression is the most common GLM for count data, and because it is robust to over-dispersion and other variances in the data (Winkelmann, 2008), we conducted our analysis using Poisson regression. Later we will show that the results of using Poisson regression on our data set are virtually identical to the results of using Negative Binomial regression.

3.3 Generalized Linear Regression Models

Regression analysis is a common tool used to explain the variance in a dependent variable, y , using independent variables, x . The most common form of regression analysis, Ordinary Least Squares, is not appropriate for our model for a number of reasons. First, OLS assumes that the residuals are normally distributed, which is not true in our case. Second, OLS is not suitable for count data because count data must take on positive integer values (Winkelmann, 2008). For this reason, count data is generally evaluated using GLMs with a link function, such as a logarithm, which forces the variables to be positive. The Poisson Regression Model is an example of a GLM; a comparison of the Poisson regression to OLS regression is shown below.

Ordinary Least Squares Regression

$$\hat{y} = \beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

Where \hat{y} = the dependent variable, x_i = the independent variables, and β_i =the effect of the independent variables on the dependent variable.

Poisson Regression

$$\hat{\lambda} = \exp(\beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k) = e^{\beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k}$$

Where $\hat{\lambda}$ = the dependent variable, x_i = the independent variables, β_i =the effect of the independent variables on the dependent variable, and e = the natural log (2.718...).

The Negative Binomial Regression Model is a GLM like the Poisson Regression Model, and also uses a log link function.

GLMs are also different from OLS models in that they have no equation to determine the effect of independent variables on the dependent variable. Instead, they rely on Maximum Likelihood Estimations, in which parameters for the model are determined so as to maximize the probability that the observed data was generated by the given model (Winkelmann, 2008). Maximum likelihood estimations are computed by mathematical algorithms employed by statistical software programs.

3.3.1 Goodness of Fit Tests for Generalized Linear Regression Models

One final difference between OLS regression and GLM regression is what methods are used to measure how well the regression equation fits the actual data, or how effectively the independent variables predict the value of the dependent variable. OLS regression uses R^2 to determine the level of fit, which is a measure of the percentage of variance in the dependent variable that is attributable to the variation in the independent variable. Because GLMs are computed using Maximum Likelihood Estimators, they do not use the OLS R^2 to determine fit, but instead use other tools that provide a measure of fit that is similar to what R^2 is to OLS. The most common measure of fit for GLMs, used for its robustness, is the McFadden R^2 (Veall & Zimmermann, 1996); we will use this measure to assess the fit of our regression models.

3.4 Inputs for Simulation

One purpose of our research was to help RailCo understand the tradeoffs between the number of extra board employees, the number of qualifications of each employee, the amount of overtime required, and total labor costs. To understand the relationships between these variables we used Monte Carlo Simulation. In Monte Carlo Simulation, a number of inputs that each obeys a specific probability distribution are defined and used to find a solution deterministically. By

running many iterations of this simulation you can gain an understanding of the best solution to the overall problem (Metropolis, N.; Ulam, S. 1949). In our case with RailCo, the inputs that obey a probability distribution are the number of absences on a given shift, the number of qualifications of incumbent employees, and the number of qualifications of extra board employees. After defining these inputs we assigned dispatchers to positions by using an optimization solver that was designed for our type of assignment problem. By changing qualification levels and the number of extra board employees and running simulation iterations we were able to investigate the impact of varying the number of extra board employees and qualifications on costs. We will now explain the specific inputs and the optimization portion of our simulation model.

3.4.1 Distributions of Employee Qualifications

The first step in building a simulation model was to determine the distribution of qualifications by day/shift of incumbent employees and the distribution of qualifications by day/shift of extra board employees. To do this, we used three relational database tables provided by RailCo, which are “Current Assignments,” “All Qualifications,” and “Qualifications by Day/Shift.” Sample data from these three tables are shown in Figure 8.

Current Assignments

Employee	Position
1103	6264
1104	6436
1105	6151
1106	6192
1358	6267
1359	6333
1360	6225
1361	6149
1364	6215

All Qualifications

Employee	Qualification
1103	6146
1103	6147
1103	6148
1103	6150
1103	6151
1103	6152
1103	6154
1103	6156
1103	6177

Qualification by Day/Shift

Qualification	DOW	Shift
6072	SUN	1
6072	MON	1
6072	TUE	1
6072	WED	1
6072	THU	1
6072	FRI	1
6072	SAT	1

Figure 8: Samples from data tables

The Current Assignments table indicates each employee’s number and the position for which he or she is the incumbent, the All Qualifications table lists every position that each employee is currently qualified to work on, and the Qualifications by Day/Shift table shows the corresponding day and shift of each qualification. Some qualifications are used during all seven days of the week, others are used only for five days, and a few are used on only two days.

Using the data in from these tables we created a distribution of employee qualifications for each day/shift combination. The distributions for extra board employees are shown in Figure 9, below.

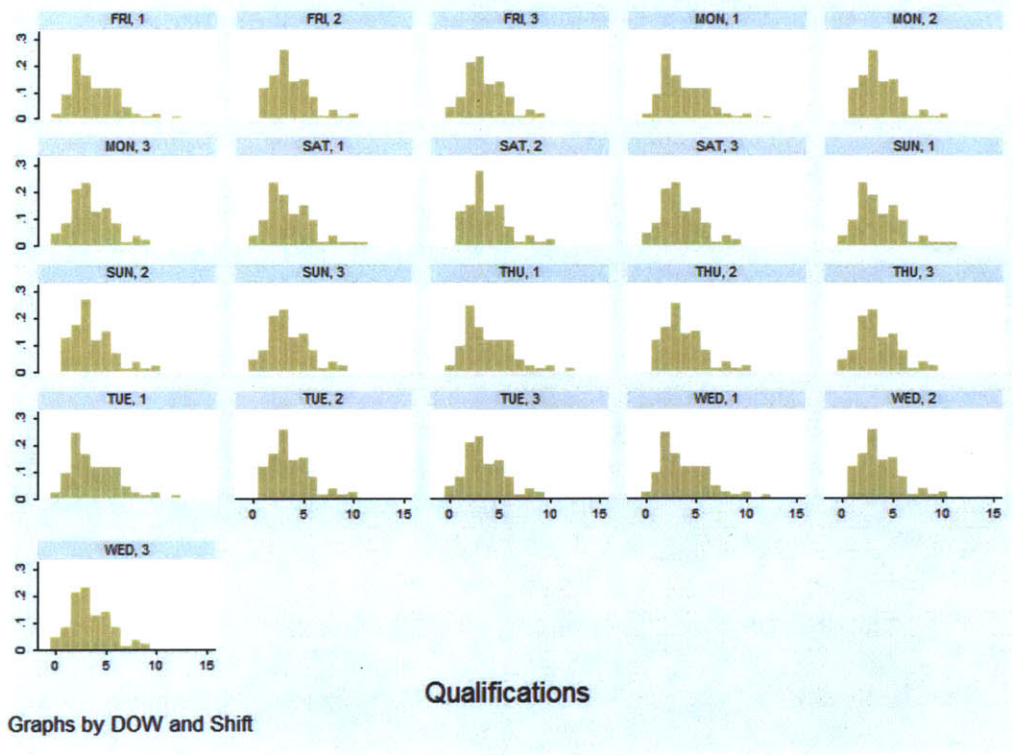


Figure 9: Distributions of qualifications for extra board employees

As is evident from Figure 9, each day/shift produces a similar distribution; we fit each of these distributions to a Negative Binomial distribution using a maximum likelihood estimator and tested the degree of fit using the Pearson Chi Square test. The statistics that describe each day/shift distribution and its level of fit to a Negative Binomial distribution is shown in Table 4.

Table 4: Statistics for distributions of extra board qualifications

		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1st Shift	Average	3.847	3.847	3.847	3.847	3.847	3.647	3.647
	Overdispersion	1.431	1.431	1.431	1.431	1.431	1.324	1.324
	Chi Square	11.750	11.750	11.750	11.750	11.750	9.045	9.045
	P-Value	0.896	0.896	0.896	0.896	0.896	0.959	0.959
2nd Shift	Average	3.753	3.753	3.753	3.753	3.753	3.694	3.658
	Overdispersion	1.107	1.107	1.107	1.107	1.107	1.122	1.145
	Chi Square	14.110	14.110	14.110	14.110	14.110	17.090	16.560
	P-value	0.722	0.722	0.722	0.722	0.722	0.517	0.554
3rd Shift	Average	3.506	3.506	3.506	3.506	3.506	3.506	3.506
	Overdispersion	1.158	1.158	1.158	1.158	1.158	1.158	1.158
	Chi Square	7.946	7.946	7.946	7.946	7.946	7.946	7.946
	P-Value	0.968	0.968	0.968	0.968	0.968	0.968	0.968

We see in Table 4 that the average number of qualifications that each extra board employee has on Monday first shift is 3.847. The over-dispersion is 1.431, and the Chi Square value for the Monday first shift distribution is 11.75, which, given sixteen degrees of freedom, produces a p-value of .896, from which we can conclude that the number of qualifications for extra board employees on Monday first shift can be appropriately modeled by a Negative Binomial distribution. Indeed, every day/shift combination produces a p-value above .05, and consequently can be modeled effectively by a Negative Binomial distribution.

We repeated this process of creating distributions for each day/shift for incumbent employees with a few minor variations. Unlike extra board employees, who can fill any position on any day without incurring extra cost, incumbent employees can only be scheduled on the unique day and shift to which they are assigned. For this reason, the incumbent's qualifications were only counted on the days and shift when the incumbent was normally scheduled. The distributions for incumbent employee qualifications are shown in Figure 10, below.

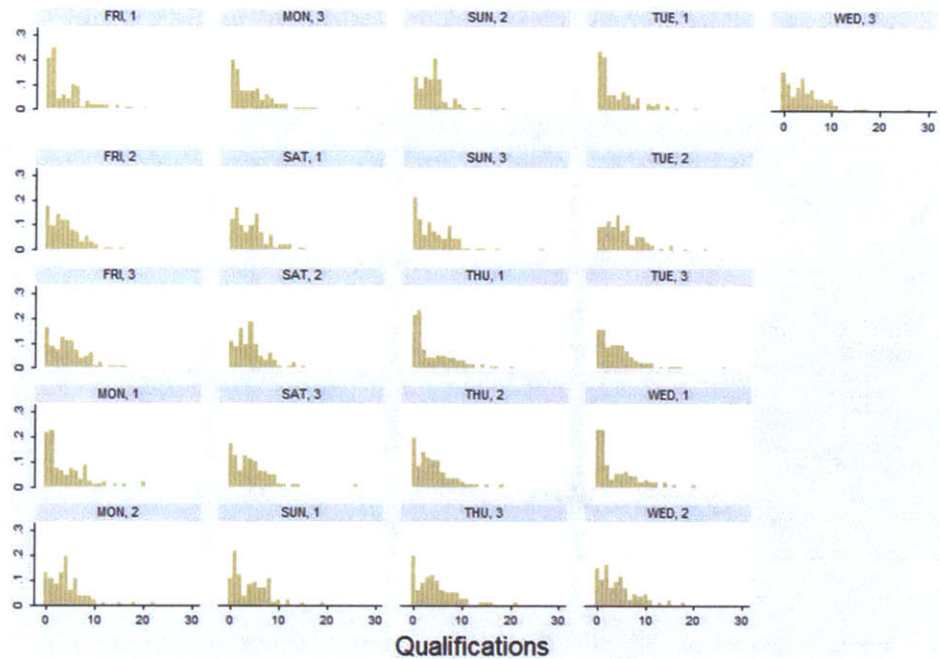


Figure 10: Distributions of qualifications for incumbent employees

As seen in Figure 10, the distributions of qualifications of incumbent employees by day/shift have more variation than the distributions of qualifications of extra board employees. We deducted one from each incumbent employee's count of qualifications and fit the remaining number to Negative Binomial distributions. We deducted one from every incumbent's count of qualifications because incumbent employees must have the qualification for the position to which they are assigned, but we were interested in knowing how many qualifications they had in addition to their regular assignment. The statistics for incumbent employees and their level of fit to Negative Binomial distributions is contained in Table 5.

Table 5: Statistics for distributions of incumbent qualifications

		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1st Shift	Average	3.891	3.846	3.833	4.114	4.023	4.175	4.241
	Overdispersion	5.437	5.799	5.735	6.008	5.599	3.358	3.550
	Chi Square	34.600	22.370	18.159	17.096	27.303	14.110	29.830
	P-Value	0.023	0.321	0.577	0.647	0.127	0.825	0.096
2nd Shift	Average	4.313	5.036	4.341	3.918	3.821	4.150	3.813
	Overdispersion	3.346	3.515	3.900	3.898	3.302	2.678	2.613
	Chi Square	36.660	22.340	17.007	10.999	9.202	15.626	30.780
	P-value	0.026	0.616	0.711	0.924	0.970	0.740	0.043
3rd Shift	Average	4.462	4.420	4.864	4.500	4.481	4.215	4.295
	Overdispersion	5.827	4.661	4.629	4.474	3.752	4.078	5.229
	Chi Square	18.240	25.176	25.916	18.150	12.970	43.140	30.650
	P-Value	0.867	0.509	0.468	0.697	0.934	0.019	0.242

Table 5 shows only four out of twenty-one day/shift p-values (Monday first, Monday second, and Saturday third, and Sunday second) that are below .05 when fit to Negative Binomial distributions. Friday third shift has the highest p-value of .934. The sum of p-values for Friday third shift is .934+.968=1.902, which is the highest sum for any day/shift combination; we selected Friday third shift for our simulation for this reason.

3.5 Optimization

We will now explain the optimization portion of our simulation model. To find a solution that minimizes cost we used a pre-programmed optimization solver that was designed for assignment problems like ours. This optimization problem is written mathematically as:

$$\text{Min } \sum_{i=1}^{N+E+1} \sum_{j=1}^N c_{ij}x_{ij} \tag{5}$$

Subject to:

$$\sum_{j=1}^N x_{ij} \leq 1, \quad i \leq N + E \tag{6}$$

$$\sum_{i=1}^{N+E+1} x_{ij} = 1 \quad j \leq N \tag{7}$$

$$x_{ij} \leq a_{ij} \tag{8}$$

$$a_{ij}, x_{ij} \in (0,1) \tag{9}$$

Where c_{ij} is the cost of assigning person i to job j ; $x_{ij} = 1$ if person i is assigned to job j at a certain time and 0 otherwise; $a_{ij} = 1$ if person i is qualified for job j and 0 otherwise. N is both the number of positions on any given day and shift, and the number of incumbent employees; E is the number of extra board dispatchers.

In solving this problem, we generated two matrices, the first of which is shown in Figure 11.

		Position					
		1	2	3	4	...	N
Incumbent Employee	1	1	1	0	0	...	0
	2	0	1	0	1	...	1
	3	0	0	1	0	...	0
	4	1	0	0	1	...	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	N	1	0	1	1	...	1
Extra Board	N+1	1	0	0	0	...	0
	N+2	0	0	1	1	...	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	N+E	1	0	0	1	...	0
Employee from Home	N+E+1	1	1	1	1	...	1

Figure 11: Qualification Matrix

In Figure 11, the 1's represent the qualifications of each employee. The horizontal numbers from 1 to N represent positions and the vertical numbers, going from 1 to N, represent individual employees. For example, incumbent employee 2 is qualified for position 2 but not qualified for

position 1. The diagonal of the matrix is filled with 1's, which represents the incumbent employees in their regular positions.

Rows from N+1 to N+E represent the extra board employees; the 1's and 0's in these rows also represent qualifications. For example, extra board employee N+2 is qualified on position 3 and 4, but not qualified on positions 1 and 2, etc.

The last row represents the pool of dispatchers that can be called from home and paid overtime to work in any position. These are all 1's because we are assuming any position on any given day and shift can be filled by an employee that can be called from home.

The second matrix is the cost matrix, shown in Figure 12.

		Position					
		1	2	3	4	...	N
Incumbent Employee	1	0	0.5	X	X	...	X
	2	X	0	X	0.5	...	0.5
	3	X	X	0	X	...	X
	4	0.5	X	X	0	...	X
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	N	0.5	0	0.5	0.5	...	0
Extra Board	N+1	0	X	X	X	...	X
	N+2	X	X	0	0	...	X
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	N+E	0	X	X	0	...	X
Employee from Home	N+E+1	1.5	1.5	1.5	1.5	1.5	1.5

Figure 12: Cost Matrix

The matrix in Figure 12 has an identical structure to Figure 11; the horizontal numbers from 1 to N represent positions and the vertical numbers from 1 to N represent employees. Each cell in this

matrix represents the cost of assigning the employee of that row to the position of that column. An “X” in a cell indicates that employee is not qualified for that position and cannot be assigned there. For example, if employee 2 is assigned to position 4, the overtime cost will be 0.5, but if that employee is assigned to position 2, which is her scheduled position, there is no extra cost; employee 4 cannot be assigned to position 1 or 3, and therefore the corresponding cells are filled with X’s. In the actual simulation program, cells with X’s will be assigned a large number that will force the optimization software not to pick any of those cells.

In Figure 12, the rows from N+1 to N+E represent extra board employees, and according to the company policy, there is no overtime cost to assign an extra board employee to a position that he/she is qualified for, and so the cells in these rows will be either X or 0.

The last row in Figure 12 represents the pool of dispatchers that can be called from home and paid overtime to work in any position. According to RailCo policy, employees that are called from home must be paid time and a half per day.

Now that we have established a matrix of qualifications and a matrix of costs, we can use the solver to find the optimal solution. As the solver is employed it will generate a third matrix that represents the assignments that lead to a minimum cost. An example of what this will look like is shown in Figure 13.

		Position					
		1	2	3	4	...	N
Incumbent Employee	1	1	0	0	0	...	0
	2	0	0	0	0	...	0
	3	0	0	1	0	...	0
	4	0	0	0	0	...	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	N	0	0	0	0	...	1
Extra Board	N+1	0	1	0	0	...	0
	N+2	0	0	0	0	...	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	N+E	0	0	0	0	...	0
Employee from Home	N+E+1	0	0	0	1	0	0

Figure 13: Solution Matrix

In this solution matrix, 1's represent the assignment of employees to positions. For example, positions 1 and 3 are filled by the incumbent employee, position 2 is filled by extra board employee N+1, and position 4 is filled by an employee called from home. The sum of the values in each column must be 1, meaning that every position must be filled by exactly one person, and the sum of each row that represents incumbent employees and extra-board employees must be 1 or 0, indicating that each person can only be assigned to a maximum of one position. The last row, which represents employees called from home, can have a sum of zero, one, or any number greater than one up to N, meaning that any number of employees could be called from home to fill vacant positions.

3.6 Simulation

We will now give a summary of the simulation using all the parameters described above. The steps to the final simulation are as follows:

1. Generate a qualification matrix with $N + E + 1$ rows and N columns, each cell having values of 0's or 1's, with 1 representing the corresponding employee qualification and 0 signifying that employee is not qualified on that position. Each incumbent employee has the qualification for his/her regular position plus some number of qualifications that is generated using the parameters of the Negative Binomial distribution for any chosen day/shift that is shown in Table 5. Each extra board employee is assigned some number of qualifications that is generated using the parameters of the Negative Binomial distribution for any chosen day/shift shown in Table 4. For both incumbent and extra board employees, the positions for which they receive qualifications are randomly distributed, with each position being equally likely. The last row of the assignment matrix is filled with 1's because any position can be filled by someone called from home.
2. Generate a cost matrix of identical size to the qualification matrix, in which each entry in the matrix represents the cost associated with that employee working in that position.
3. Generate a certain number of absences based on the parameters of the Negative Binomial distribution for unplanned absences that is shown in Table 3, and randomly assign those absences to incumbent and extra board employees, with every employee having an equally likely chance of being absent. To indicate an absence in our simulation we included a constraint in the matrix, namely, that the sum of the row for an absent employee must be zero.
4. Use a linear program solver to make assignments for all positions in a way that minimizes the total cost that results from the number of slides that occurred and the number of employees that were called from home.

5. Run 10,000 iterations of this simulation in which each iteration generates a new qualification matrix, a new set of absences, a new solution matrix, and a resulting total cost. This simulation will produce an expected cost given the pre-determined number of extra board employees and the pre-determined average number of qualifications.
6. Adjust the number of extra-board employees and the average number of the qualifications and repeat the simulation starting with step 1.

This simulation could be used to simulate any combination of day/shift, number of extra board employees, and average number of qualifications. We chose to model Friday third shift in our analysis; the corresponding “N” for our model was eighty-five, as seen in Table 6, which shows the number of regular positions that must be filled for each day and shift.

Table 6: Number of positions for each day and shift

	MON			TUE			WED			THU			FRI			SAT			SUN		
Shift	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Number of Positions	96	90	85	96	90	85	96	90	85	96	90	85	96	90	85	88	87	85	88	87	85

In our simulations we varied the number of extra board employees from zero to twenty, and used eight different qualification levels for extra board employees. The average qualification level of incumbent employees was not varied during our simulations. The average number of qualifications of extra board employees and incumbent employees is shown in Table 7.

Table 7: Mean and standard deviation of each employee type

Mean		Standard Deviation	
Normal	Extra	Normal	Extra
12.101	11.224	10.196	6.636

The average number of qualifications of extra board employees is 11.22, while on each day and shift their average number of qualifications is 3.6756. Therefore, if every extra board employee receives an average of one more qualification, the average number of qualifications per shift increases by 0.327, which is 11.22 divided by 3.6756. In our simulation we will add or deduct a multiple of 0.327 to increase or decrease of the average level of qualifications per employee.

Table 8 shows the eight qualification levels we used in our simulation and the overall qualification levels they represent.

Table 8: Average qualifications per employee

Average Qualifications per employee	Average Qualifications per employee per shift
8.22	2.525
9.22	2.852
10.22	3.179
11.22	3.506
12.22	3.833
13.22	4.160
14.22	4.487
15.22	4.814

4 Data Analysis: Regression

Our analysis aimed to construct a model that would allow RailCo to predict the number of absences that would occur on any given shift. To test whether such a model was possible, we identified several factors that might influence the number of absences on a company-wide scale. We compiled this list of possible factors after considering previous work in the literature and suggestions by RailCo scheduling management. Table 9 shows the list of factors considered in the construction of our model.

Table 9: Possible factors that influence unplanned absences

Day of the Week
Day of the Month
Month
Shift
Holidays
Football Games
Hunting Season
Snow Storms
Planned Absences

We evaluated these parameters using backward stepwise regression, in which all the parameters were included in the model, and then insignificant parameters were eliminated one at a time until only statistically significant factors remained. The null hypothesis in this case is that the effect of the parameters is zero, and the alternative hypothesis is that the effect is non-zero. Parameters with p-values above .05 were not considered to provide enough evidence to reject the null hypothesis (i.e., they were not considered to contribute significantly to unplanned absences). These factors included day of the month, day of the week, planned absences, hunting season, and football games. The parameters that produced p-values less than .05 were assumed to exhibit evidence sufficient to reject the null hypothesis of non-significance. The parameters considered significant included shift, month, selected holidays, and snow storms. The next section describes how we constructed and evaluated the parameters contained in the model. We will conclude our data analysis by examining two trends that are apparent over the four year period that we evaluated, which are the number of absences by year and the number of absences attributable to specific employees.

4.1 Statistically Significant Parameters

The four parameters that produced p-values of less than .05 were month, shift, holidays, and snowstorms.

4.1.1 Month

From our regression analysis it is apparent that month does have an impact on the number of unplanned absences at RailCo. July has the lowest average of unplanned absences so it was taken as the base value to which the other months were compared. Table 10 gives a breakdown of each month and its respective p-value.

Table 10: Regression statistics for month

Month	Coef.	Actual Effect	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
jan	0.218790	0.584938	0.049603	4.41	<.0001	0.121571	0.316010
feb	0.251001	0.671053	0.050121	5.01	<.0001	0.152766	0.349235
mar	0.234261	0.626298	0.046907	4.99	<.0001	0.142325	0.326196
apr	0.245267	0.655723	0.047538	5.16	<.0001	0.152094	0.338440
may	0.092407	0.247051	0.048202	1.92	0.055	-0.002067	0.186881
jun	0.092801	0.248105	0.047902	1.94	0.053	-0.001085	0.186687
aug	0.072701	0.194366	0.047977	1.52	0.130	-0.021333	0.166734
sep	0.003406	0.009106	0.049479	0.07	0.945	-0.093571	0.100383
oct	0.185871	0.496927	0.046714	3.98	<.0001	0.094312	0.277429
nov	0.043576	0.116500	0.049186	0.89	0.376	-0.052827	0.139978
dec	0.120119	0.321140	0.048946	2.45	0.014	0.024188	0.216050

As seen in Table 10, January, February, March, April, October and December all display p-values below the .05 level, from which we conclude that these months do have a non-zero effect on unplanned absences. May, June, August, September, and November do not produce values that are statistically different from July, the lowest month. There may be many factors that contribute to this difference of absences among months, such as the weather, seasonal variations

in sickness among the workforce, the number of allotted planned absences, or other idiosyncratic factors not evaluated in our model.

4.1.2 Shift

Shift also has a non-zero effect on the number of unplanned absences that occurred over the four year period. RailCo can expect a slightly higher number of absences on second and third shifts as compared to first shifts. These findings are in Table 11. It makes intuitive sense that third shift would have the highest number of unplanned absences given that it starts sometime between the hours of 9:00 PM and 4:59 AM.

Table 11: Regression statistics for shift

Shift	Coef.	Actual Effect	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
shift2	0.066634	0.178148	0.026224	2.54	0.011	0.015237	0.118032
shift3	0.101265	0.270733	0.023696	4.27	<.0001	0.054822	0.147709

4.1.3 Holidays

The Holidays we evaluated in our regression model were selected based on a 2010 study conducted by Worldatwork. This study found that the average number of paid holidays in the United States is 8.7. This study also presented a list of the most common paid holidays, from which we took the top nine, which are displayed in Table 12.

Table 12: Nine most common paid Holidays

New Years Day
Thanksgiving
Labor Day
Memorial Day
Independence Day
Christmas Day
Day after Thanksgiving
Christmas Eve
Presidents Day

For New Year’s Day we combined the third shift of New Year’s Eve because it technically runs into New Year’s Day. In addition to these nine paid holidays, we created one other holiday parameter called “Federal Holidays,” which is a combination of Martin Luther King Jr. Day, Columbus Day, and Veterans Day. These three days are Federal Holidays but were not included in the top nine Holidays cited in the Worldatwork study.

The results, shown below in Table 13, indicate that the most popular Holidays have the somewhat surprising effect of lowering the number of absences for any given day. This is true for New Year’s, President’s Day, Independence Day, Thanksgiving, Christmas Eve, and Christmas. The remaining holidays, Labor Day, Memorial Day, the Friday after Thanksgiving, and the combination of Federal Holidays, are not significant.

Table 13: Regression statistics for Holidays

Holiday	Coef.	Actual Effect	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
newyears	-0.722219	-1.930722	0.209295	-3.45	0.001	-1.132429	-0.312009
presidents	-0.420272	-1.122878	0.206649	-2.03	0.042	-0.825297	-0.015248
memorial	-0.418113	-1.115345	0.226170	-1.85	0.065	-0.861397	0.025172
independence	-0.916658	-2.448851	0.303559	-3.02	0.003	-1.511622	-0.321694
labor	-0.295194	0.000000	0.221066	-1.34	0.182	-0.728476	0.138088
thanksgiving	-1.171696	-3.104133	0.335387	-3.49	<.0001	-1.829043	-0.514350
thanksgivingfriday	-0.330449	0.000000	0.221151	-1.49	0.135	-0.763897	0.103000
christmaseve	-0.841878	-2.248154	0.260941	-3.23	0.001	-1.353313	-0.330443
christmas	-0.762535	-2.035175	0.252826	-3.02	0.003	-1.258065	-0.267006
federal	0.010323	0.000000	0.101771	0.10	0.919	-0.189144	0.209790

Based on the values in Table 13, RailCo can expect two to three fewer absences on shifts on popular holidays. While these results may be somewhat counter-intuitive, it seems reasonable that the stigma of calling in sick on a holiday is great enough to give dispatchers extra motivation to arrive at work. It may also be that employees are extra-motivated to come to work on holidays because of the negative effect it has on other employees that are forced to fill in for the absent employee.

4.1.4 Snowstorms

One final parameter that shows a high degree of significance in contributing to unplanned absences is snow storms. Twenty shifts included in our model were noted as having had significant snowstorms; data on snow storm dates and severity was taken from list of snow events compiled by the National Weather Service for the local area. The effect of snowstorms on absences is shown in Table 14. According to our analysis, RailCo can expect an extra 2.16 absences on shifts that have snowstorms.

Table 14: Regression statistics for snowstorms

Event	Coef.	Actual Effect	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
snow	0.810674	2.167346	0.090367	8.97	<.0001	0.633558	0.987790

This finding is not particularly helpful in predicting future absences in the long range because it is difficult to accurately forecast the weather, but it may be helpful for RailCo in their short term planning.

4.1.5 Planned Absences

There were three types of planned absences included in this parameter: scheduled vacation, float vacation, and personal days. These planned absences did not have a significant effect on unplanned absences by themselves, as seen in Table 15. However, removing them from the model it has the effect of lowering the statistical significance of other parameters and reducing the overall predictive capability of the model. This effect of parameters being influenced by each other is known as co-linearity. For this reason, it is useful to consider planned absences as a significant factor even though their p-value, at .106, is slightly higher than .05.

Table 15: Regression statistics for planned absences

Parameter	Coef.	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
plannedabsences	-0.0132658	0.0081977	-1.62	0.106	-0.0293329	0.0028014

4.2 Statistically Insignificant Factors

There are a number of parameters that do not give us enough evidence to conclude that their effect on absences is not zero. These include football games, day of the month, and day of the week. We will now discuss these parameters in detail.

4.2.1 Hunting Season

Hunting season in the local area, begins every year on the first Sunday in November and ends on the first Sunday in January. For the purposes of our analysis, we assigned the “beginning of hunting season” parameter to the first two days of hunting season every year, and the “end of

hunting season” parameter to the last two days of hunting season each year. As seen in Table 16, the beginning of hunting season was found to be insignificant regarding absences, with a p-value of .493, and the end of hunting season was insignificant with a p-value of .08. The data do not provide sufficient evidence to conclude that the effect of hunting season in November and December is greater than zero.

Table 16: Regression statistics for hunting season

Parameter	Coef.	Actual Effect	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
Beg Hunt Season	-0.096593	0.000000	0.140805	-0.69	0.493	-0.372566	0.179380
End Hunt Season	0.217245	0.580145	0.124163	1.75	0.080	-0.026110	0.460601

4.2.2 Football Games

We evaluated two types of football games: the Super Bowl and regular season NFL games. The “NFL” parameter in Table 17 includes all dates on which NFL games were played, and the “Super Bowl” parameter includes all shifts of the days on which the Super Bowl was played. From the p-values of .702 and .284 for NFL and Super Bowl, respectively, there is insufficient evidence to conclude that NFL football games or the Super Bowl affect unplanned absences.

Table 17: Regression statistics for football games

Parameter	Coef.	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
NFL	0.01894	0.04944	0.38	0.702	-0.077954	0.115830
Super Bowl	-0.19899	0.18556	-1.07	0.284	-0.562688	0.164712

4.2.3 Day of the week and Day of the Month

The data for day of the month is shown in Table 18. The first day of each month was used as the base number to which all the other days of the month were compared. There are a few days of the month that would be significant at the .05 level, but given that there are thirty-one days of the

month to be considered, we used a .01 test of probability for this parameter. At this level of significance there were no days of the month that showed significance.

Table 18: Regression statistics for days of the month

Day	Coef.	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
2	-0.09403	0.07144	-1.32	0.188	-0.234052	0.045990
3	-0.12347	0.07218	-1.71	0.087	-0.264952	0.018007
4	-0.04346	0.07184	-0.61	0.545	-0.184265	0.097338
5	-0.09289	0.07223	-1.29	0.198	-0.234458	0.048668
6	-0.15652	0.07365	-2.13	0.034	-0.300872	-0.012172
7	-0.04398	0.07161	-0.61	0.539	-0.184337	0.096377
8	-0.03191	0.07106	-0.45	0.653	-0.171180	0.107370
9	-0.00183	0.07017	-0.03	0.979	-0.139356	0.135698
10	-0.04430	0.07120	-0.62	0.534	-0.183860	0.095251
11	-0.09151	0.07222	-1.27	0.205	-0.233062	0.050052
12	-0.12982	0.07246	-1.79	0.073	-0.271836	0.012203
13	-0.12053	0.07256	-1.66	0.097	-0.262744	0.021689
14	-0.01421	0.07073	-0.2	0.841	-0.152836	0.124410
15	-0.14450	0.07323	-1.97	0.048	-0.288020	-0.000977
16	-0.11024	0.07263	-1.52	0.129	-0.252593	0.032120
17	-0.09610	0.07218	-1.33	0.183	-0.237561	0.045369
18	-0.06988	0.07167	-0.97	0.330	-0.210356	0.070605
19	-0.12545	0.07265	-1.73	0.084	-0.267853	0.016944
20	-0.02127	0.07078	-0.3	0.764	-0.159993	0.117461
21	-0.08126	0.07183	-1.13	0.258	-0.222041	0.059530
22	-0.02187	0.07105	-0.31	0.758	-0.161118	0.117374
23	-0.09835	0.07206	-1.36	0.172	-0.239597	0.042891
24	-0.09169	0.07406	-1.24	0.216	-0.236854	0.053473
25	-0.08073	0.07422	-1.09	0.277	-0.226193	0.064734
26	-0.09976	0.07264	-1.37	0.170	-0.242138	0.042616
27	-0.14174	0.07316	-1.94	0.053	-0.285130	0.001650
28	-0.14704	0.07324	-2.01	0.045	-0.290588	-0.003494
29	-0.08360	0.07325	-1.14	0.254	-0.227158	0.059957
30	-0.09013	0.07413	-1.22	0.224	-0.235413	0.055155
31	-0.03626	0.08357	-0.43	0.664	-0.200044	0.127532

Day of the week also showed no statistical significance when evaluated at the .05 level. Table 19 details the p-values for each day of the week using Monday as the base.

Table 19: Regression statistics for day of the week

Day	Coef.	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
Tue	-0.02933	0.03581	-0.82	0.413	-0.099510	0.040857
Wed	0.00673	0.03549	0.19	0.850	-0.062832	0.076288
Thu	0.04165	0.03540	1.18	0.239	-0.027730	0.111033
Fri	0.01388	0.03560	0.39	0.697	-0.055904	0.083655
Sat	0.01165	0.03566	0.33	0.744	-0.058237	0.081533
Sun	0.04180	0.03736	1.12	0.263	-0.031413	0.115019

4.3 Marginal Effects

Table 20 is a summary of all the parameters that we have discussed that show statistically significant impact on the number of absences. Any parameters not listed in this table did not provide enough evidence to indicate significance at the .05 level. As a reminder, the actual effect of any given parameter will vary from day to day, so the numbers listed here are the average effect over the range evaluated. For example, in the month of April, RailCo can expect to have an average of .57 more absences than the base of July on any given shift, and .17 more absences on second shifts as compared to first shifts.

Table 20: Summary of actual effects of statistically significant parameters

Parameter	Avg. Effect	Std. Err.	z	P>z	Lower 95% int	Upper 95% int
jan	0.58494	0.13272	4.41	0.000	0.32481	0.84507
feb	0.67105	0.13414	5.00	0.000	0.40814	0.93396
mar	0.62630	0.12554	4.99	0.000	0.38025	0.87235
apr	0.65572	0.12724	5.15	0.000	0.40634	0.90510
oct	0.49693	0.12498	3.98	0.000	0.25198	0.74187
dec	0.32114	0.13089	2.45	0.014	0.06460	0.57768
shift2	0.17815	0.07013	2.54	0.011	0.04070	0.31560
shift3	0.27073	0.06340	4.27	0.000	0.14647	0.39500
snow	2.16735	0.24243	8.94	0.000	1.69220	2.64249
newyears	-1.93072	0.55983	-3.45	0.001	-3.02797	-0.83348
presidents	-1.12288	0.55258	-2.03	0.042	-2.20591	-0.03985
independence	-2.44885	0.81188	-3.02	0.003	-4.04011	-0.85759
thanksgiving	-3.10413	0.89692	-3.46	0.001	-4.86207	-1.34620
christmas	-2.03518	0.67619	-3.01	0.003	-3.36048	-0.70988
christmaseve	-2.24815	0.69794	-3.22	0.001	-3.61609	-0.88022

4.4 Significance of Model

Even though many parameters are statistically significant in their effects on unplanned absences, the overall model is not a reliable predictor of the number of absences that will occur on any given shift. The final model has a McFadden Pseudo R^2 value of only .0179. For perspective, an OLS model with a value of .0179 would mean that less than 2% of the variance in the dependent variable is attributable to variations in the independent variables. The McFadden Pseudo R^2 value of .0179 is not an exact percentage like the OLS R^2 , but it has similarly low predictive ability. In short, the model is useful to give a general idea of a few factors that contribute to unplanned absences, but it will not produce accurate predictions for unplanned absences. Unplanned absences occur in a random pattern and there does not appear to be a satisfactory manner to accurately predict their fluctuation.

4.5 Goodness of Fit for Negative Binomial Regression

Because the Negative Binomial distribution was a slightly better fit for the distribution of unplanned absences, we compared the results of Negative Binomial regression to our Poisson regression results. The list of parameters that were statistically significant was identical, and the Pseudo R^2 actually decreased when using a Negative Binomial regression. Table 21 compares the coefficients of the parameters between the Negative Binomial and Poisson regression.

Table 21: Comparison of results from Poisson and Negative Binomial regression

<u>Poisson regression</u>			<u>Negative binomial regression</u>		
Pseudo R2: 0.0183			Pseudo R2: 0.0170		
Parameter	Coef.	P>z	Parameter	Coef.	P>z
jan	0.21911	0.000	jan	0.21879	0.000
feb	0.25073	0.000	feb	0.25100	0.000
mar	0.23460	0.000	mar	0.23426	0.000
apr	0.24529	0.000	apr	0.24527	0.000
may	0.09234	0.060	may	0.09241	0.055
jun	0.09287	0.057	jun	0.09280	0.053
aug	0.07279	0.136	aug	0.07270	0.130
sep	0.00347	0.945	sep	0.00341	0.945
oct	0.18615	0.000	oct	0.18587	0.000
nov	0.04372	0.382	nov	0.04358	0.376
dec	0.12040	0.016	dec	0.12012	0.014
shift2	0.06681	0.012	shift2	0.06663	0.011
shift3	0.10138	0.000	shift3	0.10127	0.000
plannedabsences	-0.00496	0.112	plannedabsences	-0.00496	0.106
snow	0.81160	0.000	snow	0.81067	0.000
endhuntingseason	0.21598	0.089	endhuntingseason	0.21700	0.080
newyears	-0.72183	0.001	newyears	-0.72217	0.001
presidents	-0.41983	0.045	presidents	-0.42000	0.042
memorial	-0.41736	0.068	memorial	-0.41718	0.065
independence	-0.91596	0.003	independence	-0.91597	0.003
thanksgiving	-1.16113	0.001	thanksgiving	-1.16107	0.001
christmaseve	-0.84171	0.001	christmaseve	-0.84090	0.001
christmas	-0.76296	0.003	christmas	-0.76124	0.003

4.6 Other Considerations: Yearly Trends and Absences by Employee

It is important to note that there is a large difference in unplanned absences by year. Year 2012 had a smaller number of absences when compared to every other year, as seen in Figure 14.

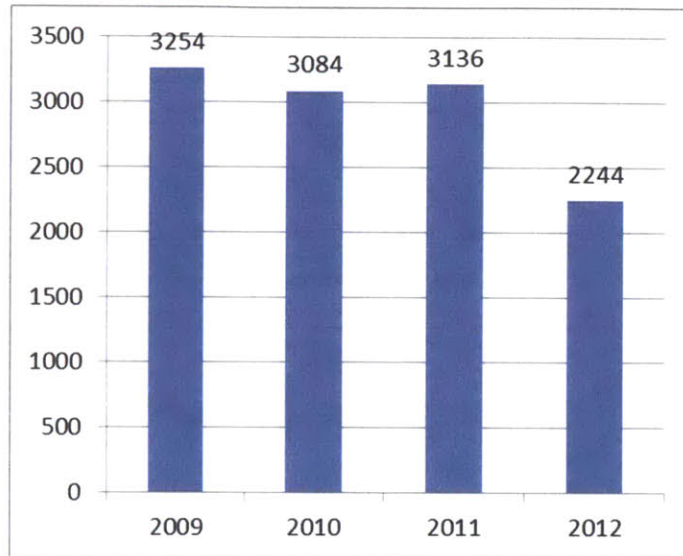


Figure 14: Total absences by year

According to RailCo, the total number of employees from 2009 to 2012 did not change, however, there were slight union agreement modifications that may have contributed to this marked decrease in 2012 unplanned absences. An additional insight is illustrated in Figure 15, which shows the cumulative absences from 2009 to 2012. The more gradual slope of the 2012 curve indicates not only that there were fewer absences that year but also that there were fewer employees that had high numbers of absences.

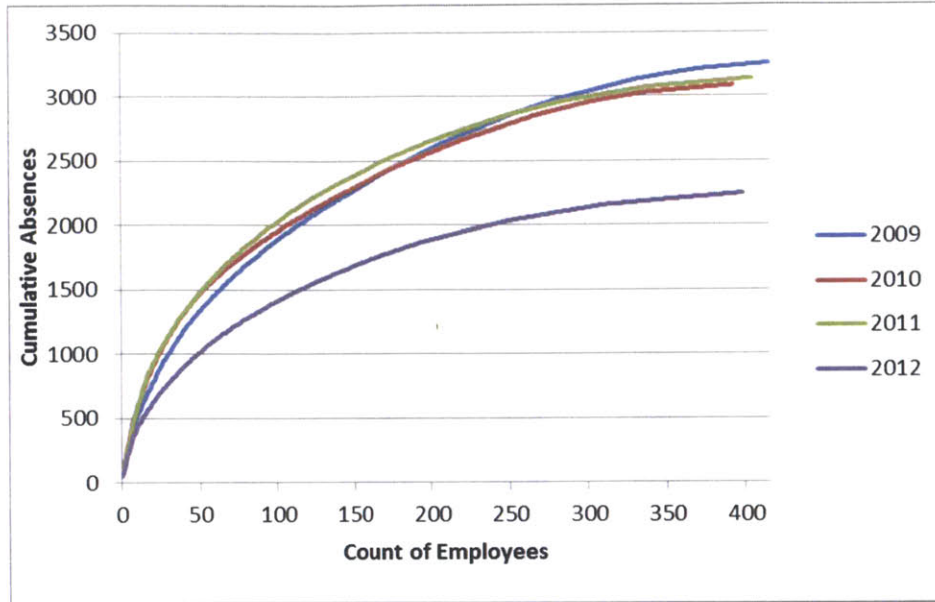


Figure 15: Cumulative absences by year

5 Data Analysis: Simulation

Our purpose in this section is to investigate the relationship among extra board size, average qualification level of extra board employees, and labor costs; this will form the basis for RailCo’s hiring and training strategies. In our analysis we made the assumption that RailCo is able to fill all incumbent planned vacations with extra board employees without sliding employees or calling employees from home. It should be noted that planned vacation vacancies vary by day and shift; the average number of planned vacations by day and shift is shown below in Table 22. Our analysis does not take into account any cost associated with assigning extra board employees to the planned vacations represented in this table.

Table 22: Average planned absences

	1st Shift	2nd Shift	3rd Shift
Sun	10.56	6.30	10.34
Mon	14.77	5.17	9.83
Tue	15.66	6.32	9.67
Wed	14.40	8.11	10.50
Thu	12.32	8.08	10.75
Fri	13.89	7.87	7.98
Sat	10.39	7.31	8.21

One Key Performance Indicator that RailCo uses in their scheduling management is the percent overtime. This is calculated as the extra cost divided by the cost of filling each position with an incumbent employee. For example, if over the course of a year there were 5,000 slides and 500 employees called from home, and we assume 90 positions on each of 3 shifts, 365 days a year, then the percent overtime would be $(.5*5000 + 1.5*500)/365*3*90 = 3.3\%$. In our discussion of costs, we will use the following terms as defined below:

- “Slide cost” is defined as the total amount paid to slide incumbent employees from their regularly assigned position. Under union rules, RailCo must pay an additional .5 of employee’s normal daily salary for every slide occurrence.
- “Overtime cost” is defined as the cost of calling dispatchers from home to fill a position. Under union rules RailCo must pay employees called from home 1.5 times his/her normal daily salary.
- “Extra cost” is the combined total of slide cost and overtime cost.

The section below explains our analysis of each of these costs change based on varying levels of extra board employees and the average number of qualifications of extra board employees. As stated earlier in section 3.6, we included extra board size from zero to twenty, and eight different

qualification levels, from three below the current average to four above. We used Friday third shift in our simulation as explained in section 3.4. From Table 4 we see the current average number of qualifications of extra board employees for Friday third shift is 3.506.

5.1 Slide Cost

Figure 16, below, illustrates the change in slide cost with different sizes of extra board and varying averages of qualifications. The vertical axis represents the cost of the slides, which is one half the number of slides. The horizontal axis is the number of extra employees above those used to fill planned vacation days, and each line shows a different level of qualifications.

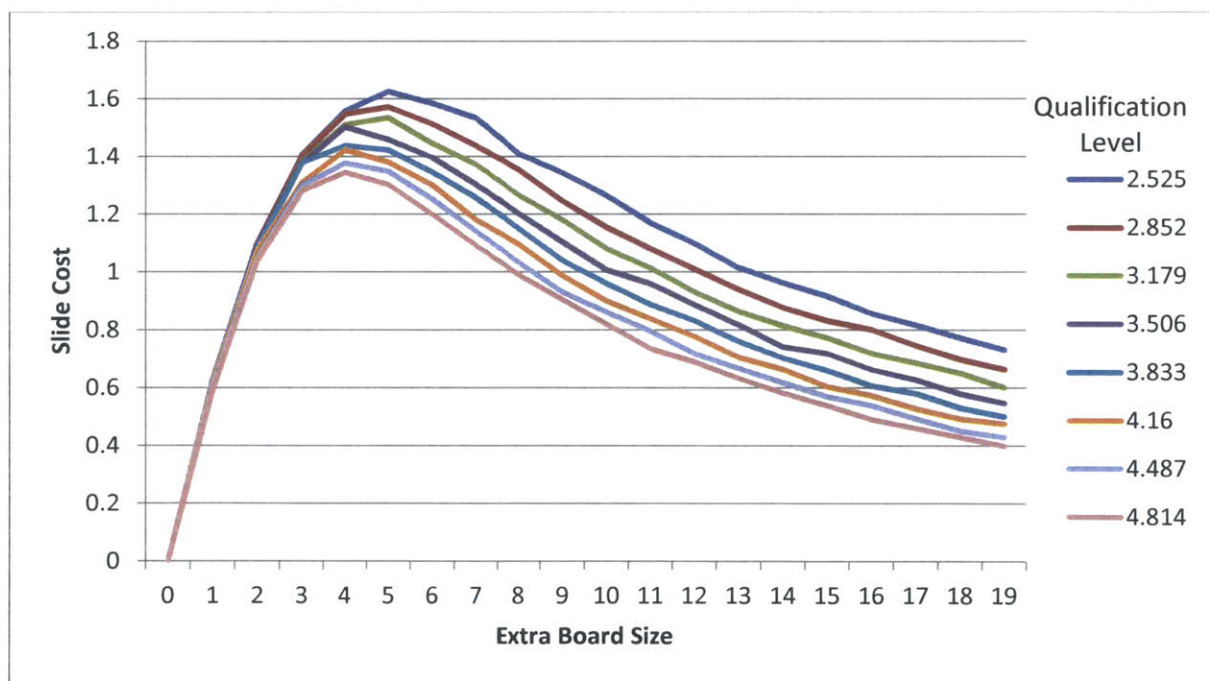


Figure 16: Slide cost

We can see when we begin to add people on extra board from 0, the slides cost increase sharply. This is because the maximum number of slides is limited to the number of extra board employees. The slide cost reaches its peak when the extra board size is about five for any

qualification level we tested, after which it decreases gradually because the more extra board employees we have, the more likely it is that we can fill the absence directly with the extra board. We can also see slide costs are slightly higher when the qualification level is lower. This is because if employees have more qualifications, they will be more likely to have the qualifications needed to fill each position.

5.2 Overtime Cost

Figure 17, below, shows overtime cost, or the cost of calling employees from home.

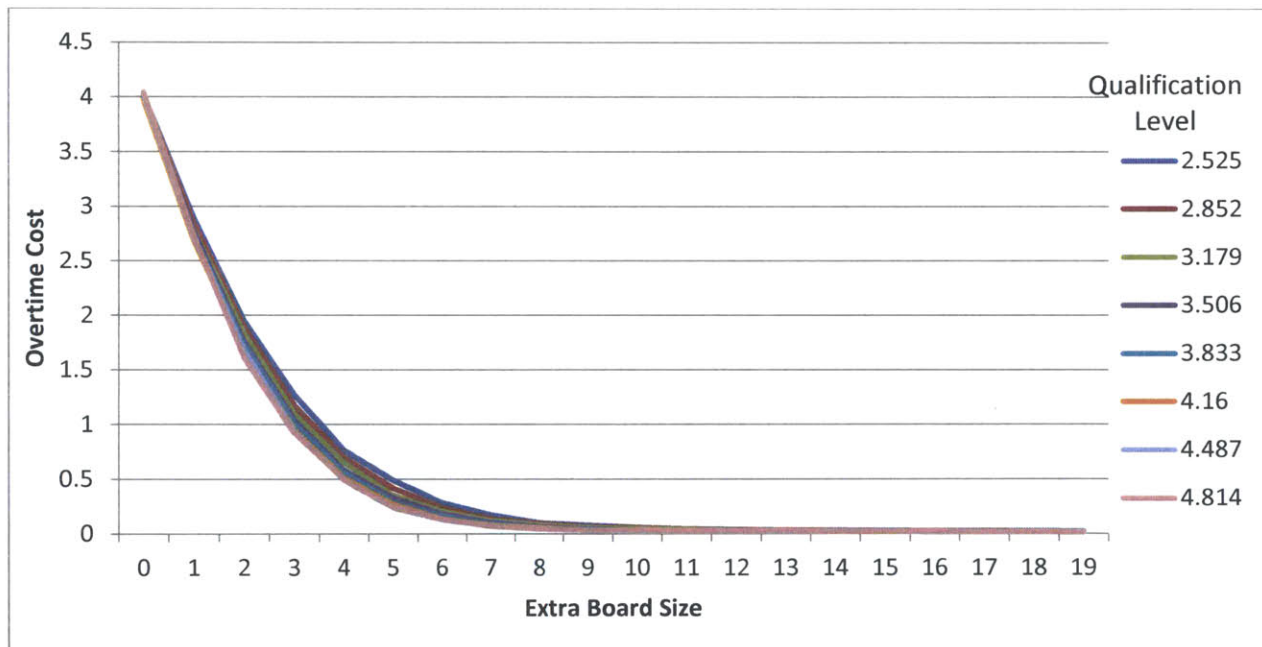


Figure 17: Overtime cost

The total overtime cost drops sharply as we add extra board employees. The fact that all the lines representing the different qualification levels are close to each other indicates that varying the number of qualifications does not make a large impact on the number of employees that must

be called from home. (It is important to note that this is based on the assumption that additional qualifications are randomly assigned.)

Another illustration of this relationship is shown in Figure 18, below. For lower numbers of extra board employees, total costs gradually decrease as qualifications increase. As the number of extra board employees increases, the qualification level does not play an important role in overtime cost (i.e., the slope of the line becomes flat).

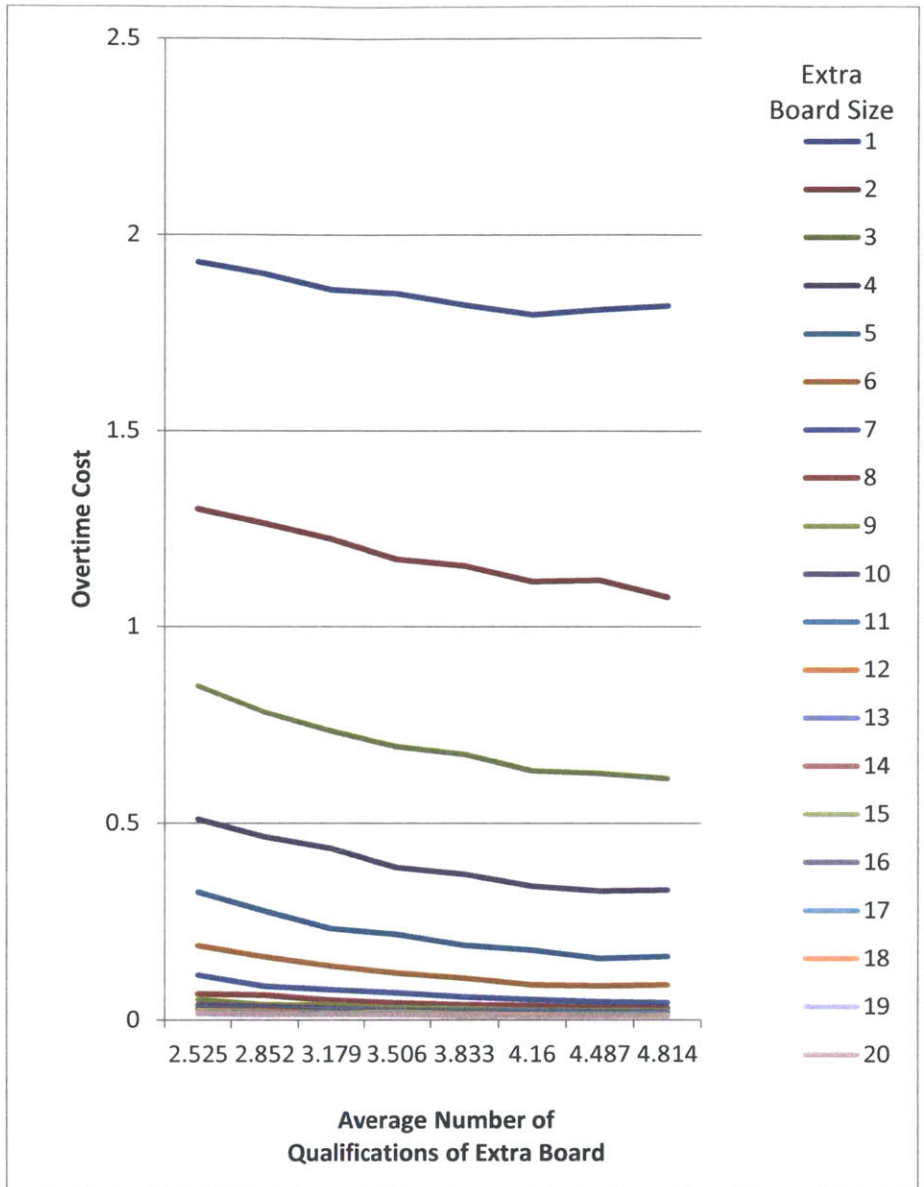


Figure 18: Overtime cost by qualifications

5.3 Extra Cost

Figure 19, below, shows the total extra cost as we add more people to the extra board for each qualification level.

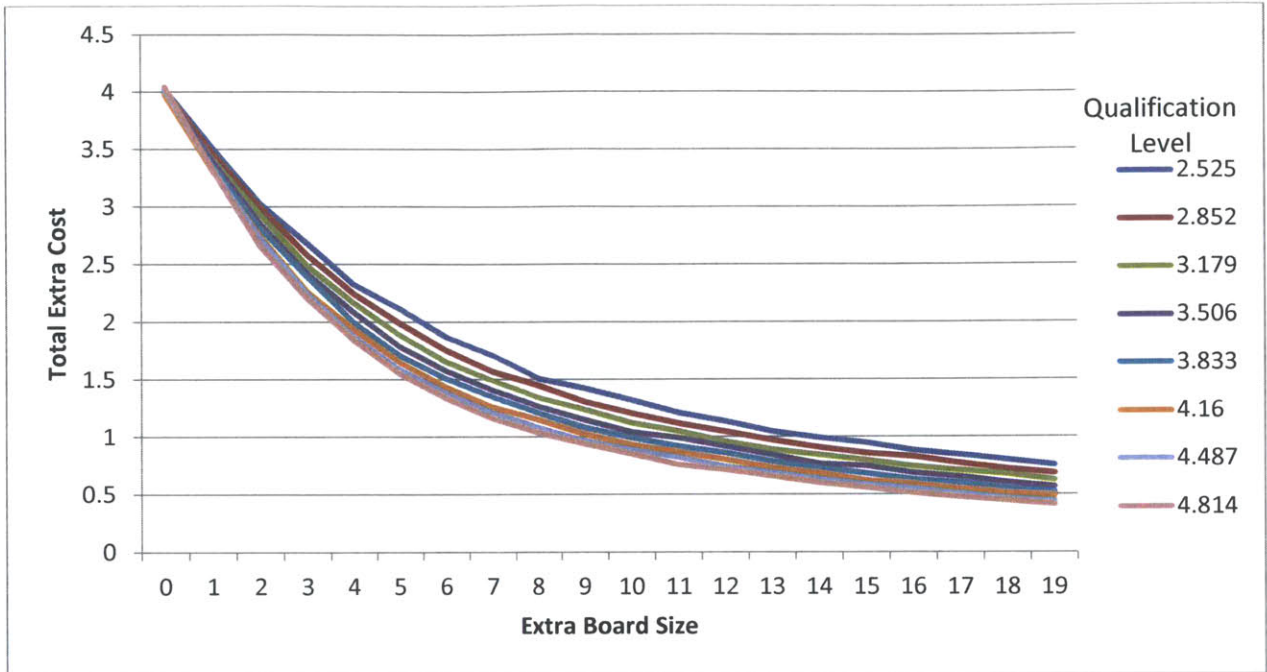


Figure 19: Extra cost

Extra cost always decreases when there are more people on extra board, and the effect is significant, especially when the extra board size is smaller.

The relationship between slide cost and overtime cost as the extra board increases is shown in Figure 20.

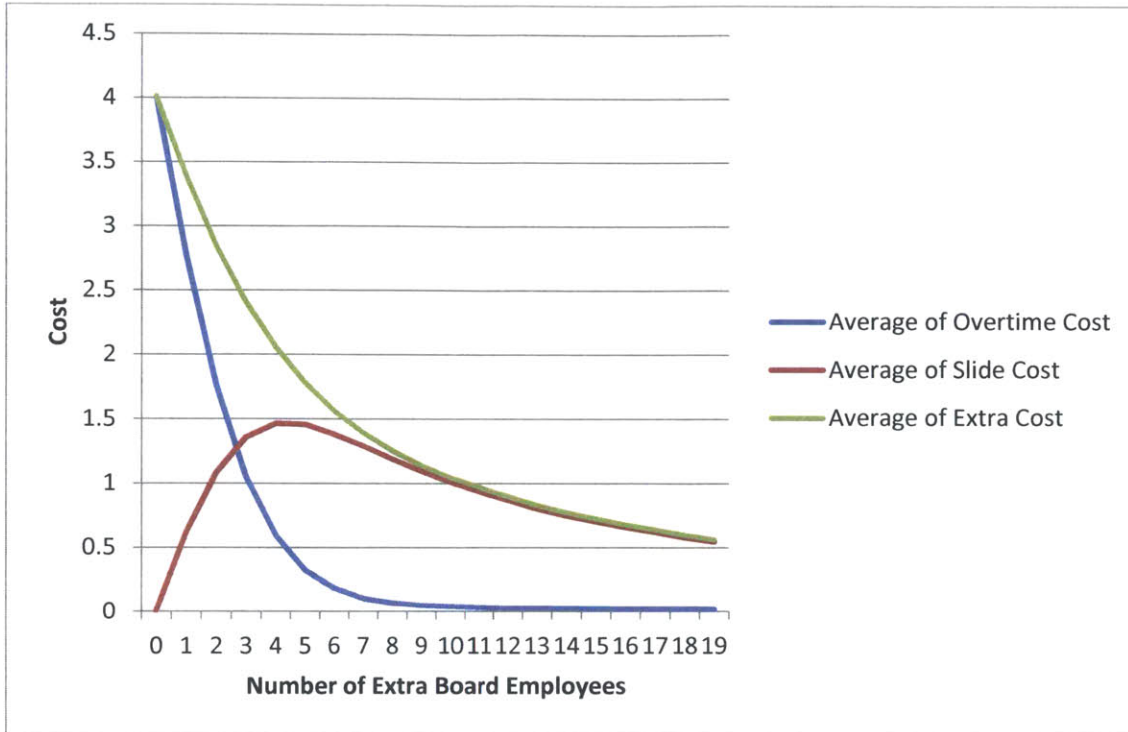


Figure 20: Slide cost, overtime cost, and extra cost

5.4 Current Cost

We will now evaluate RailCo’s current extra board size and number of qualifications. Currently, RailCo has eighty-five extra board employees and each of them work on five shifts a week. By summing the data in Table 22 we found that the average number of planned absences in a week is 208, so the number of shifts that extra board employees work above what is needed to fill planned absences is $85 \times 5 - 208 = 217$. Given that there are 21 shifts every week, the average number of extra board employees per shift after accounting for planned absences is $217 / 21 = 10.33$. We also know from Table 8 we know that the average number of qualifications of extra board employees is 11.22. To illustrate the change of costs associated with increasing extra board employees or qualifications we will use 10 as the number of extra board employees,

and 11 as the average number of qualifications for extra board employees. Figure 21 shows the current situation given those numbers.

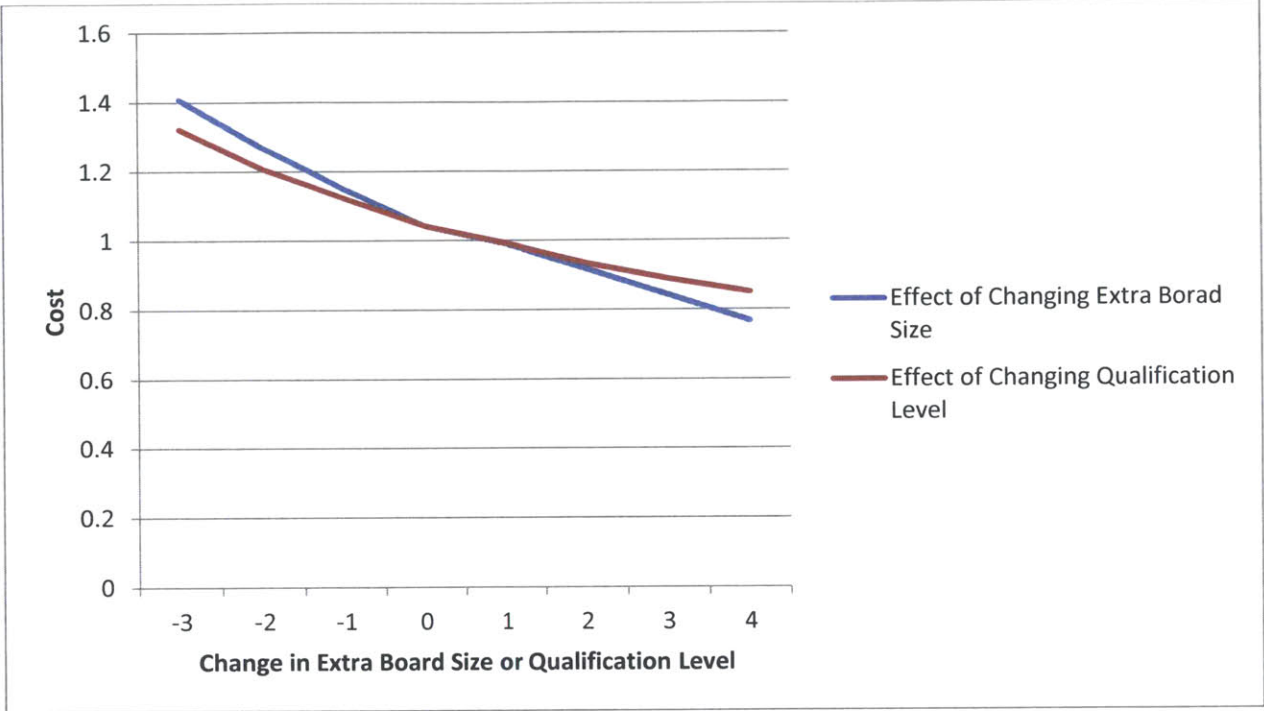


Figure 21: Effect of changing extra board size or qualification

In Figure 21, point 0 on the horizontal axis represents the current situation, and moving along the graph to the right or left shows the cost change from increasing or decreasing the number of extra board employees or qualifications.

6 Conclusion

The purpose of our research was to determine what factors, if any, could be used to predict unplanned dispatcher absences. Through Poisson regression analysis, we have shown that there are several factors that do influence the number of absences that occur on any given shift. The factors that show statistical evidence of increasing the expected number of absences are

snowstorms, second and third shifts (when compared to first shift), and the months of January, February, March, April, and December (using July as the base month). The factors that show statistical evidence of decreasing the number of expected absences are holidays including New Year's Day, President's Day, Independence Day, Thanksgiving, Christmas Eve, and Christmas. The factors that do not provide statistical evidence of affecting the number of unplanned absences are day of the month, day of the week, football games, and hunting season.

Even though we successfully identified factors that influence the number of unplanned absences, the Mcfadden R^2 value of our final regression model was only .0174, which suggests that the combination of all the variables studied do not lead to a satisfactory way to predict expected absences. However, we have shown that unplanned absences by shift can be effectively modeled with a Negative Binomial distribution.

Through Monte Carlo simulation we have shown that extra costs decrease as the number of extra board employees increases. This decrease is greater when the number of extra board employees is lower. Overtime costs decrease sharply as the number of extra board employees increases, and slide costs are highest when there are approximately five extra board employees above the number needed to fill all planned absences, and then gradually decrease as extra board size increases. Furthermore, increasing the number of qualifications of each extra board employee makes a small impact on decreasing slide and overtime costs; this impact becomes almost negligible when the size of the extra board is large.

6.1 Other Considerations in Employee Staffing

While our analysis has focused on slide and overtime costs, RailCo may also consider the total labor costs associated with increasing extra board employees or the average number of

qualifications. RailCo can decrease slide and overtime costs by hiring extra board employees, but they must pay each of these employees a full salary even if they do not have an assignment every day. We can define “total labor cost” as the total cost above what RailCo would pay in an ideal scenario where every incumbent worked their regular position and there was no need for extra board employees, slides, or overtime. Figure 22, below, shows total labor cost as the number of extra board employees increases.

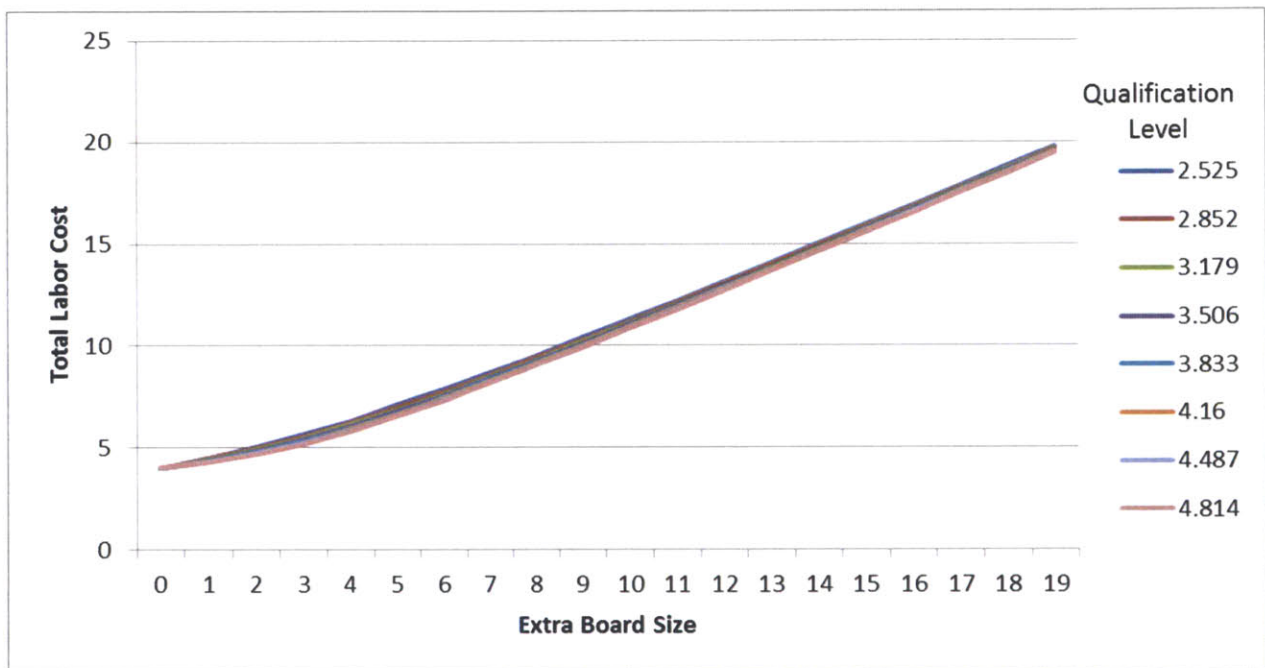


Figure 22: Total labor cost

Figure 21 shows that regardless of qualification level, it will always cost more overall to increase the size of the extra board. Figure 23 shows the cost of staffing extra board employees, total labor costs, and extra costs.

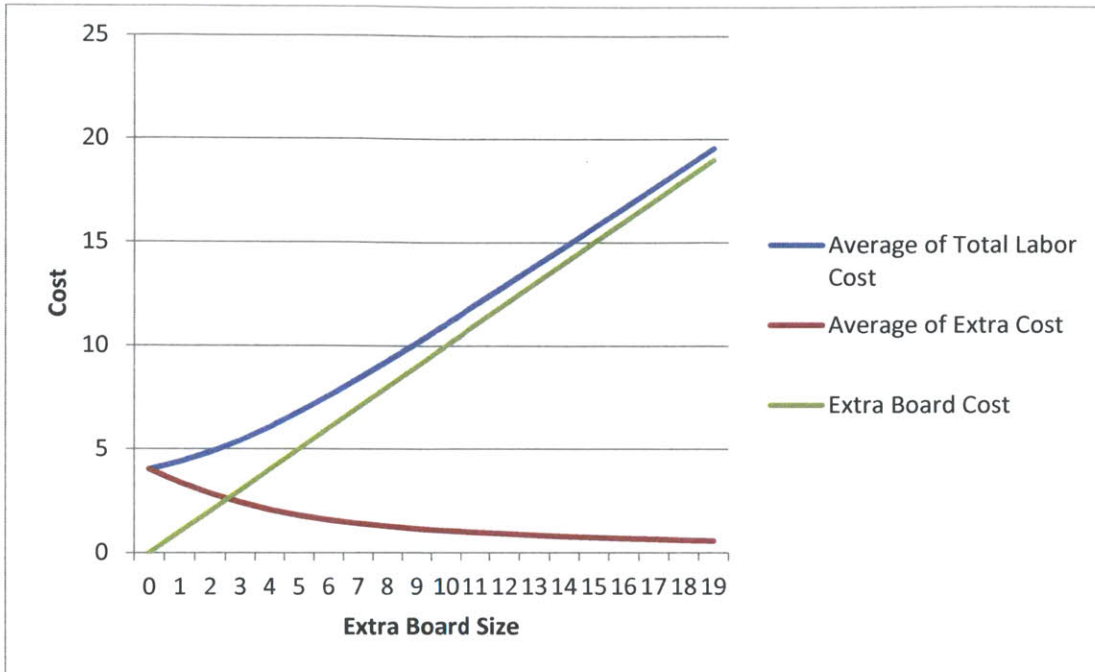


Figure 23: Total labor cost, extra cost, and extra board cost

As seen in Figure 23, adding extra board employees will decrease slide and overtime costs, but the amount of decrease is much less than the added labor costs from having more employees.

Even though the lowest total labor cost will be achieved when there are just enough extra board employees to fill planned absences, but not more than that, there are several other factors that

RailCo must take into account in their hiring strategy. For example, a smaller extra board will make it necessary to call people from home more often, which may have adverse effects on

employee morale and labor union relations. RailCo may also need to consider that if they have a

lower number of employees there may have rare occasions when they cannot fill a position with a qualified employee from home because of labor rules.

6.2 Future Research

Further areas of research for RailCo may include using factors that influence employee absences on an individual scale, such as age and job satisfaction, to aid them in predicting absences, although the costs associated with aggregating this kind of data are likely to exceed the benefits. This thesis was also only concerned with one company in one geographical location, so further research examining different companies in different locations may not produce similar results because of cultural, geographic, and demographic differences. RailCo may also explore the impact of variability in absences on costs, and the impact on costs of one additional absence.

RailCo may also want to investigate the cost associated with training employees, and how that would change their ideal mix of qualifications and extra board size. RailCo may also investigate the importance of non-cost factors in their staffing strategy, such as employee morale and union agreements. Understanding these relationships will allow them to build an optimal staffing strategy based on their unique parameters.

7 Bibliography

Bureau of Labor Statistics. (2011). Household Data Annual Averages. Table 46: Absences from work of employed full-time wage and salary workers by age, sex, race, and Hispanic or Latino ethnicity. Washington DC: Superintendent of Documents, U.S. Government Printing Office.

Bureau of Labor Statistics. (2012). Work Absences due to bad weather: analysis of data from 1977 to 2010. Washington DC: Superintendent of Documents, U.S. Government Printing Office.

Carpenter, G., & Wyman, O. (June 2010) Survey on the Total Financial Impact of Employee Absences. Retrieved from <http://www.mercer.com/press-releases/1383785>

Chaudhury, M., & Ng, I. (1992). Absenteeism predictors: Least squares, rank regression, and model selection results. *Canadian Journal of Economics*, 25(3), 615.

De Boer, E. M., Bakker, A. B., Syroit, J. E., & Schaufeli, W. B. (2002). Unfairness at work as a predictor of absenteeism. *Journal of Organizational Behavior*, 23(2), 181-197.

Egerváry, E. (1931). On combinatorial properties of matrices, *Mat. Lapok*, 38, 16-28.

Harrison, D. A., & Martocchio, J. J. (1998). Time for absenteeism: A 20-year review of origins, offshoots, and outcomes. *Journal of Management*, 24(3), 305-350.

Hausknecht, J. P., Hiller, N. J., & Vance, R. J. (2008). Work-unit absenteeism: Effects of satisfaction, commitment, labor market conditions, and time. *Academy of Management Journal*, 51(6), 1223-1245.

Hultberg, T. H., & Cardoso, D. M. (1997). The teacher assignment problem: A special case of the fixed charge transportation problem. *European journal of operational research*, 101(3), 463-473.

Konig, D. (1936). *Theorie der endlichen und unendlichen Graphen*, Leipzig: Akad. Verlagsges. mbH.

Kuhn, H. W. (1955), The Hungarian method for the assignment problem. *Naval Research Logistics*, 2: 83-97. doi: 10.1002/nav.3800020109

- Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247), 335-341.
- Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial & Applied Mathematics*, 5(1), 32-38.
- Prater, T., & Smith, K. (2011). Underlying factors contributing to presenteeism and absenteeism. *Journal of Business & Economics Research*, 9(6), 1-14.
- VEALL, M. R. and ZIMMERMANN, K. F. (1996), PSEUDO-R2 MEASURES FOR SOME COMMON LIMITED DEPENDENT VARIABLE MODELS. *Journal of Economic Surveys*, 10: 241–259. doi: 10.1111/j.1467-6419.1996.tb00013.x
- Weaver, R. (2010, June 8). Cost of Presenteeism Surpasses Absenteeism. Retrieved from <http://www.examiner.com/human-capital-in-detroit/cost-of-presenteeism-surpasses-absenteeism>
- Winkelmann, R. (2008). *Econometric Analysis of Count Data 5th Edition*. Berlin Germany: Springer-Verlag Berlin Heidelberg
- WorldatWork, 2010. *Paid Time Off Programs and Practices*, retrieved May 15, 2013 from <http://www.worldatwork.org/waw/adimLink?id=38913>