

# Data Mining and Decision Support for Operational Problems Solutions in the Railroad Industry

Pavel Brusilovskiy and Robert Hernandez  
Service Design and Planning, Conrail Corporation

## Abstract

One of the major problems in the railroad industry is high yard elapsed time (YET)<sup>1</sup> [1]. The following dependent variables reflect the condition of YET problem situation: freight car elapse time and probability of inbound car making its outbound train connection. There are many input variables that could influence YET such as inbound and outbound train performance, train structure, yard activities, in particular decision maker preferences and yard capacity. This paper describes the data mining process to identify crucial factors for an unstructured problem, like YET reduction and usage of the identified factors in decision making. Reliable crucial factors can be obtained through the combination of verification-driven and discovery-driven data mining processes. The framework of data mining tools consist of SAS/BASE, SAS/STAT, SAS/QC, SAS/GRAPH, SAS/OR, SAS/IML (treedisc macro).

## Introduction

There has been much publicity about the capabilities of data mining to exploit large amounts of data to improve competitive business intelligence. Most business managers should consider seriously on initiating data mining projects to solve some of their business problems. However, many business managers that deal with semi-structured and unstructured problems might not recognize the need to use data mining tools due to the supporting data mining advertisements and case studies. Most of the data mining material relates to well structured problems. The purpose of this paper is to demonstrate the potential of data mining for unstructured problems.

## Problems Structure

Complex problems can be defined as three types: structured, semi-structured, and unstructured problems: [2]

**Structured problems** are ones for which all structural elements such as goals, alternatives, criteria and environmental conditions are known, defined and understood. Typically, these types of problems have standard solution methods and techniques that are clearly understood for solving well structured problems.

**Unstructured problems** have no standard solutions for resolving the situation. Some or all of the structural elements such as goals, alternatives, criteria and environmental conditions are undefined, ill defined or unknown. For example, goals may be poorly defined, alternatives may be incomplete or non-comparable, choice criteria may be hard to measure or difficult to line to goals.

**Semi-structured problems** are between structured and unstructured problems.

## Data Mining for Structured and Unstructured Problems

Data mining process will be different for structured and unstructured problems. One of the data mining processes that is applied in the business community is SEMMA [3]. SEMMA stands for Sample, Explore, Modify, Model, and Assess. The basic purpose of any discovery data mining is to automatically uncover important information hidden in large amounts of data. A common name for this process is discovery-driven data mining [4]. Another component of data mining is verification-driven data mining process. This process allows the decision maker to express and verify organizational and personal domain knowledge and hypotheses. We realized that both types of data mining processes were necessary to achieve successful results in unstructured problems.

### Structured Problem

For well structured problems, discovery-driven data mining is appropriate process to achieve significant results. This is not the case for unstructured and semi-structured problems which require the usage of both discovery-driven data mining and verification-driven data mining. A few examples should illustrate the difference in the data mining processes for structured and unstructured problems. Many issues at credit card companies are structured problems and the use of discovery-driven data mining would be most appropriate. For example, a credit card issuing company has a number of descriptors on each customer. The credit history for each customer is known and therefore have been place in the corresponding class of good, medium, or poor credit risk. Using a decision tree technique, it revealed that if a customer's income is over \$35,000 and the age bracket is between 40 and 50, and the customers lives in Philadelphia then the customer is a good credit risk. Because this is a structured problem, the data mining results directly correspond to the decision and normally do not require any expert opinions, strong domain knowledge, incorporation of decision maker's preferences, and numerous data mining iterations.

### Unstructured Problem

We recognize that many operational problems fall into the unstructured category and one in particular is the yard elapsed time (YET) reduction problem. Referring to the credit card issuing company example, we showed the results are relatively predictable due to the structured nature of this problem. The customer demographics data should be readily available in a properly designed data warehouse. In YET study, many of the input variables were not readily available in the data warehouse and the results were unpredictable due to the unstructured nature of the

<sup>1</sup> YET is the length of time a freight car remains idle in a classification yard before moving to its next destination.

problem. Railroad experts believed that the key factor to reduce YET was to improve on-time train performance for late arriving trains. Another belief by the experts was that freight cars on early arriving trains had a higher likelihood of making their scheduled outbound train connection than late arriving trains. This mental picture created an assumption over the years that freight cars that made their scheduled outbound train connection would typically have lower YETs than freight cars that miss their scheduled outbound train connection. Our initial test found a weak relationship between YET and the two input variables mentioned above: early and late arriving trains. The two prevailing assumptions were not true and to find any possible relationship with YET, we needed to develop new variables not readily available in the data warehouse and basically ignored or not recognized by the experts. Our final results revealed that some of the newly developed variables, which reflected unknown environmental conditions, were more significant than the train performance (known environmental conditions) issues and produced a good model for prediction. The results from the newly developed variables were completely unpredictable due to going beyond the known environmental conditions of on-time train performance. Consequently, unpredictability and unknown factors is not an issue for the credit card company example because the results have to relate to known factors of customer demographics. Unstructured problems could have unknown environmental conditions and unexpected results and therefore require that the analyst must have a strong domain knowledge of the problem under consideration, high quantitative skill level, expert opinion involvement, and use of both verification-driven data mining and discovery-driven data mining methods on iterative basis. The results from this YET study needed all these requirements and strong visualizations in order to prevent the least resistance from key decision makers on new information that contradict their dominant assumptions or paradigm.

### Yard Elapsed Time Study

There are three major operational components that operations has responsibilities which are handling and prioritizing inbound trains, yard activity management, and outbound train performance. We also did not consider yard capacity restrictions. A study to include all components of YET would not be practical in terms of time due to the complexity of the problem. To manage through the complexity, we decomposed the YET reduction problem into three steps: (1), inbound train study, (2) inbound plus outbound train study (3), inbound, outbound plus yard capacity study. The analyses in this paper correspond to inbound train study and its association with YET and cars with deviation greater than 12 hours from their scheduled YET (Figure 1).

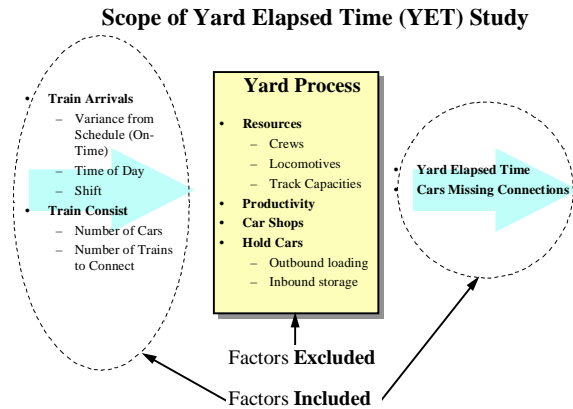


Figure 1

Some the questions we wanted to answer:

- Does inbound train performance significantly effect YET?
- Does the train structure on the inbound train significantly effect YET?
- Which trains with improved train performance would significantly reduce YET?

### Data Structure

The data structure<sup>2</sup> was constructed in the following manner:

Field Name	Field Description	Data
BSYMBOL	Train Arrival Brief Symbol	ELSE
ARACTDTE	Actual Date/Time of Train Arrival	5/7/97 7:00
AR_TIME	Actual Time of Train Arrival	7:00 AM
TRN_TIME	Actual Train Date/Time - Schedule Train Date/Time	4 hours
OT	Train arrived late or early	Late
BLOCKNUM	Number of cars on the arrival train	100
DESTNUM	Number of destinations for cars on the arrival train	15
DEVIAT	TRN_TIME Groupings	p_small
SHIFT	The crew shift when the train arrived	1
MIS_NUM	Number of mis-connected cars	24
YET	Average YET for arrival train	29 hours

Table 1

Number of Trains Analyzed: 2,394  
 Period of Observation: 5/28/97 to 6/17/97

This core data structure was used for statistical analysis purposes as well as transforming data into other variable fields, like number of cars per outbound train symbol. The objective of the constructed variables is to uncover possible combinations that are significant to YET.

<sup>2</sup> Data not included in YET analysis: Extra trains, cars with YET greater than 96 hours, and relay cars

## General Results (Railroad Level)

### Data mining results

General results were a compilation of approximately 25 data mining iterations, 100 newly constructed variables, and exploitation of many dissimilar statistical methods and models. Often times, the iteration results were discussed with railroad experts and management staff which lead to new hypothesis formulation and testing (or another iteration). The following results are related only to the inbound trains study.

#### I. Train Performance

- A. Inbound train performance does not influence YET when dataset includes both early and late arrival trains
- B. Mean values of YET are different for early and late inbound trains
- C. Late inbound train performance increases the number of cars with deviations greater than 12 hours from their scheduled YET (Only inbound trains that arrive on or after their scheduled time)

#### II. Train Structure

- A. Higher ratio of scheduled cars to total cars on inbound trains decreases YET
- B. Higher number of inbound scheduled cars per outbound train symbol decreases YET
- C. Higher number of inbound nonscheduled cars per outbound train symbol decreases YET

#### III. Other Factors

- A. Lower ratio of cars with deviations greater than 12 hours from their scheduled YET to total cars decreases YET

The next three sections will explain in more detail the findings for train structure and train performance.

#### I. Train Performance

##### A. Inbound train performance does not influence YET when dataset includes both early and late trains

There are many examples to confirm this finding and Yard A was randomly chosen to illustrate our result. Yard A has two train symbols, Train A and Train D that on average are late by 7.30 hours and 8.99 hours, respectively. Those trains are very late and still manage to achieve an equal or lower average YET relative to the yard. On the other side of inbound train performance are Train B and Train C, they arrive early by 0.728 hours and 0.137 hours, respectively - but have higher average YETs relative to the yard (See Figure 2 and 3). Consequently, train performance for all trains does not influence YET, but probably other factors do, such as late and early train service prioritization and/or scheduled and network design. Another important issue with this finding is that equal treatment of early and late inbound trains washes out relations between YET and train attributes. All Box-Whisker plots are products of SAS/QC.

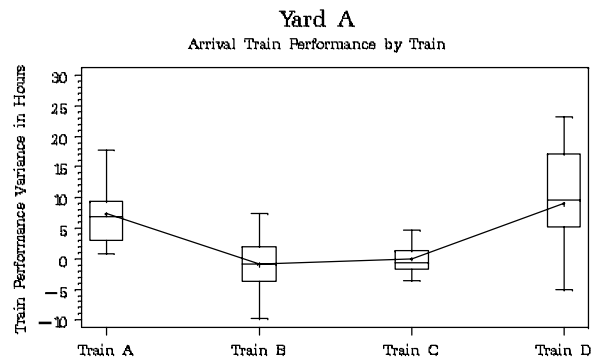


Figure 2

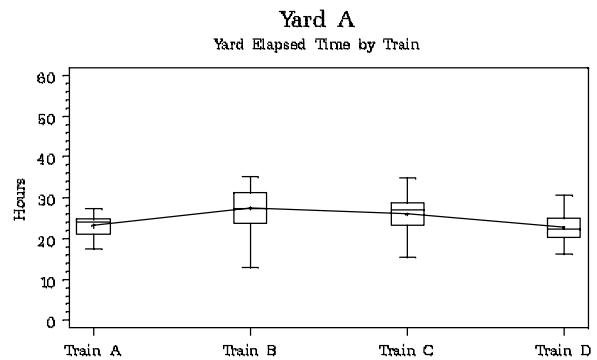


Figure 3

##### B. Mean values of YET are different for early and late trains

Train A and Train B have average early time arrivals of 3.74 hours and still their average YETs are higher than most trains that have later average train performance such as Train C and Train D. Figures 4 and 5 illustrate that early train arrivals have different YET outcomes than late arriving trains. This distinguish of early and late arriving trains having different YET's was also revealed through the decision tree results using the treedisc macro, SAS/IML and SAS/OR [5].

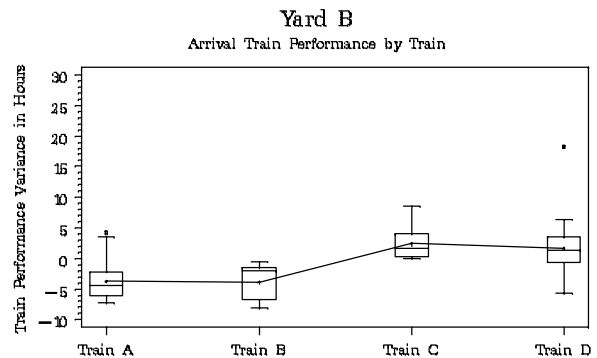


Figure 4

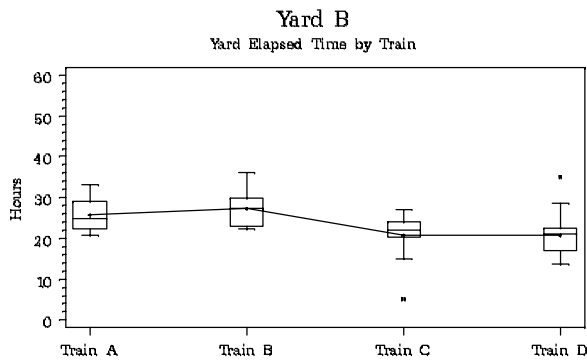


Figure 5

**C. Late inbound train performance increases the number of cars with deviations greater than 12 hours from their scheduled YET**  
(Only trains that arrive on or after their scheduled time)

The later a train arrives from schedule the higher average number of cars that will have deviations greater than 12 hours from their scheduled YET. Using Yard C as an example and studying late train time performance (actual minus schedule) in three categories: 0 to less than 1 hour, 1 hour to less than 4.5 hours, and greater or equal to 4.5 hours (See Table 2). A simple interpretation for trains in the 1 to less than 4.5 hours category is that on average these trains have YET of 21.5 hours and 11.6 cars with deviations greater than 12 hours from their scheduled YET. The total amount of trains in this category is 65. Table 2 shows that late arrival train performance has some influences on cars with deviations greater than 12 hours from their scheduled YET.

Train Performance	YET	> 12 Hours	N
0 to less than 1 hour	19.227	4.567	36
1 to less than 4.5 hours	21.544	11.631	65
Greater or equal to 4.5 hours	21.465	19.75	30

Table 2

**II. Train Structure**

**A. Higher ratio of scheduled cars to total cars on inbound trains decreases YET**

For inbound trains only, this finding has the second highest contribution to YET out of all the significant variables found in our studies. On a railroad level, the interpretation of the finding is that the higher the ratio of scheduled cars to total cars for the inbound trains the lower the average car yard elapsed time (YET). The finding also has a strong significance at a yard level, but of course, each yard is slightly different in terms of impact to YET. Figure 6 illustrates the significant difference between the YET means for scheduled and nonscheduled cars. All vertical bar charts are products of SAS/GRAPH.

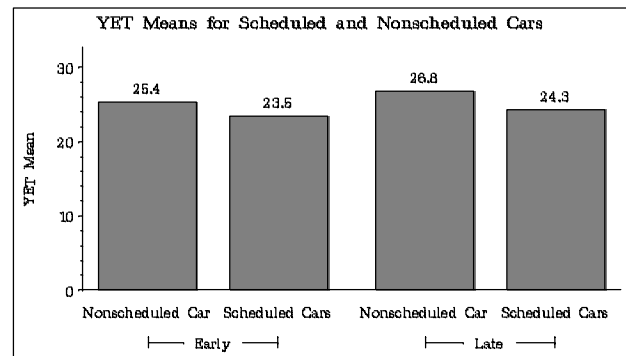


Figure 6

**B. Higher number of inbound scheduled cars per outbound train symbol decreases YET**

One of our original studies identified that number of outbound train symbols, total number of cars, and number of inbound cars per inbound train symbol were significant to YET. These findings were good but not the best for decision making because of the finding of scheduled cars to total cars on inbound trains decreases YET. We separated scheduled and unscheduled cars and found that inbound trains with higher number of scheduled cars per outbound train symbol have lower average YET (See Figure 7).

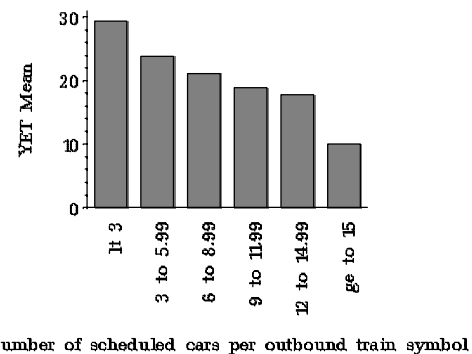
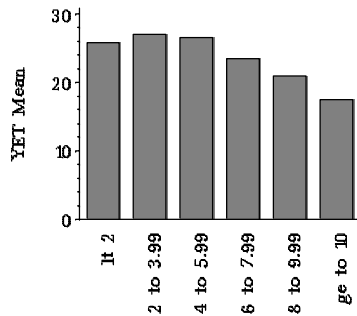


Figure 7

**C. Higher number of inbound nonscheduled cars per outbound train symbol decreases YET**

Inbound trains with higher number of unscheduled cars per outbound train symbol have lower average YET (See Figure 8 - Page 5).



Number of nonscheduled cars per outbound train symbol

Figure 8

**III. Other Factors**

**Lower ratio of cars with deviation greater than 12 hours from their scheduled YET to total cars decreases YET**

Our intent is to predict YET based on inbound trains, but YET includes more than just inbound train structure and performance. There are some strong interrelationships with inbound trains and outbound trains and how it effects YET. For example, number of cars per outbound train symbol has a strong interrelationship with the inbound side (i.e. number of inbound cars) and the outbound side (i.e. number of outbound train symbols). Ratio of cars with deviations greater than 12 hours from their scheduled YET to total cars is no different because these cars include inbound and outbound train performance as well as yard activity management. We did find a relationship with inbound train per-

formance and inbound cars with deviations greater than 12 hours from their scheduled, but a very weak one (i.e.  $R^2$  of 0.07). It means that inbound cars greater than 12 hours from their scheduled YET are influenced by other factors besides inbound train factors.

**Interaction with Service Prioritization and Train Structure Findings**

The train performance findings show the relationship (or lack of) with inbound train performance and two output variables, YET and cars with deviations greater than 12 hours from their scheduled YET. The studies revealed that late and early inbound trains are typically treated differently in terms of YET and service prioritization decision making. For example, why do we have trains that arrive extremely late and have lower YET than trains that arrive relatively on time? We don't know all the reasons for service prioritization such as yard capacity and departure train prioritization, but it is apparent that decision making is different for early and late inbound trains. Therefore, it is in our best interest to separate early and late inbound trains and create new variables based on the findings in our foundation studies.

Separating early and late inbound trains uncovered how train performance interacts with the train structure findings. There are some new variables to introduce and corresponding description at the **train level**. These findings are in ranked order based on highest contribution to YET (Table 3). Standardized coefficient values are products SAS/STAT.

**Table 3 - Findings and Descriptions**

Controllable Factor			Power of variables according to their impact YET reduction					
			Early Trains			Late Trains		
			Rank	Standardized Coefficient	Sign	Rank	Standardized Coefficient	Sign
Non-Actionable	mis_prop	ratio of inbound cars greater than 12 hours from their scheduled YET to total cars	1	0.44	+	1	0.54	+
Actionable	car_prop	ratio of scheduled inbound cars to total cars	3	0.29	-	2	0.36	-
Actionable	s_dens	number of scheduled inbound cars per outbound train symbol	2	0.35	-	3	0.24	-
Actionable	u_des	ratio of nonscheduled inbound cars per outbound train symbol	4	0.21	-	4	0.18	-
Slightly Actionable	des2prop	ratio of total unique nonscheduled outbound train symbols to total outbound train symbols	5	0.18	-	7	0.12	-
Slightly Actionable	des_prop	ratio of total unique scheduled outbound train symbols to total outbound train symbols	6	0.17	-	6	0.16	-

			Power of variables according to their impact YET reduction					
			Early Trains			Late Trains		
Controllable Factor	Variable	Description	Rank	Standardized Coefficient	Sign	Rank	Standardized Coefficient	Sign
Slightly Actionable	late_des	train performance deviation (actual- schedule) per outbound train symbol	N/A	N/A	N/A	5	0.18	-
Slightly Actionable	earl_des	train performance deviation (schedule - actual) per outbound train symbol	7	0.14	-	N/A	N/A	N/A
Actionable	late	train performance deviation (actual-schedule)	N/A	N/A	N/A	8	0.11	+
Actionable	early	train performance deviation (schedule-actual)	8	0.14	+	N/A	N/A	N/A
Non-actionable	mis_dens	number of cars greater than 12 hours per outbound train symbol	9	0.05	+	9	0.08	+

The standardized coefficient with the highest value in table 3 on page 5 and 6 would have the highest impact to YET. In this case, it would be ratio of inbound cars with deviations greater than 12 hours from their scheduled YET to total cars. We discussed in a previous section (III. Other Factors) that inbound train structure and performance have minor influences on cars with deviations greater than 12 hours from their scheduled YET. This variable is uncontrollable from an inbound train perspective and needs to be understood in future studies from an outbound train perspective. With that mind, we still have variables with a significant impact to YET. Standardized coefficients revealed that important controllable factors like car\_prop and s\_dens contribute significantly more to YET than train performance. This means that we have other controllable factors besides train performance to reduce YET.

The variables with a high impact to YET are good to know from a strategic perspective and provides us the opportunity to decide which factors will achieve the most value for our efforts. But, the unanswered question is how much will it save us in terms of reduction in cycle time? Using the best regression model and the variables from table 3 on page 5 and 6, we have predicted the average YET based on various levels of process improvements. The following matrices in tables 4 and 5 show the predicted YET reductions for early and late inbound trains at 5%, 10%, 15%, and 20% improvement. Regression model predictions are products of SAS/STAT.

**Table 4 - Early Inbound Trains - YET prediction improvement**

%of Change	Mean Improvements				YET Prediction
	mis_prop	car_prop	s_dens	Early	
0%	21%	80%	7.12	2.86	24.4
5%	20%	84%	7.47	2.71	23.5
10%	19%	88%	7.83	2.57	22.6
15%	18%	92%	8.18	2.43	21.7
20%	17%	96%	8.54	2.28	20.8

**Table 5 - Late Inbound Trains - YET improvement prediction**

%of Change	Mean Improvements				YET Prediction
	mis_prop	car_prop	s_dens	Late	
0%	27%	81%	7.65	4.83	25.3
5%	25%	85%	8.03	4.58	24.1
10%	24%	89%	8.41	4.34	22.8
15%	23%	93%	8.80	4.10	21.6
20%	21%	97%	9.18	3.86	20.3

**Summary**

To achieve improvement on all of these nine identified factors of YET would be a monumental assignment. Fortunately we have three basic contributors to YET that makes it more feasible to manage and achieve a significant reduction in YET :

1. Increase the number of scheduled cars to total cars on inbound trains
2. Increase number of inbound scheduled cars per outbound train symbol
3. Improve train performance for both early and late arriving trains

We have a high degree of confidence that if we simultaneously improve all three areas we will reduce YET by a significant amount. For example, to reduce YET from 25.3 to 22.8 (10%) for late inbound trains, we need to do the following:

- increase ratio of inbound scheduled cars to total cars from 81% to 89% (or 10% increase)
- increase number of inbound scheduled cars per outbound train symbol from 7.6 to 8.4 (or 10% increase)
- improve train performance from 4.8 to 4.3 hours (or 10% increase)

To reduce YET from 24.4 to 22.6 (7%) for early arriving trains, we need to do the following:

- increase ratio of scheduled cars to total cars on inbound trains from 80% to 88% (or 10% increase)
- increase number of inbound scheduled cars per outbound train symbol from 7.1 to 8.3 (or 10% increase)
- improve train performance from 2.8 to 2.6 hours (or 10% increase)

**YET Reduction Recommendations**

- Policy or Business Rule Issues:
  - Reducing the % of cars greater than 12 hours from their scheduled YET is a good strategy
- Service Delivery Actions to Reduce YET
  - Increase ratio of scheduled cars to total cars on inbound trains
  - Improve on-time train performance for both early and late inbound trains
- Service Design Actions to Reduce YET
  - Create inbound trains with more cars per outbound train symbol

- Multi-criteria quality assessment at the train level. [6]
- Performance Measures
  - Provide measures to Service Design & Service Delivery to quickly identify potential opportunities from the areas mentioned above
  - Examine outbound train performance to fully predict % of cars with deviations greater than 12 hours from their scheduled YET

**References**

1. Martland Carl. D., “Benchmarks For Rail Hump Yard Performance,” MIT Affiliated Lab Working Paper 94-3, pp. 1-37.
2. Raiffa, H. “ Preferences for multi-attributed alternatives”, The Rand Corporation , CA 1969.
3. SAS Institute Inc (1997). SAS Communications Second Quarter 1997, “The Jewel in the Data Mining Crown”, pp. 6-7, Cary, NC: SAS Institute Inc.
4. Cabena Peter , Pable Hadjinian, Rolf Stadler and Jaap Verhees. “Discovery Data mining from Concept to Implementation”, Prentice Hall,1997.
5. SAS Institute Inc (1995). Treedisc Macro - Beta Version, SAS Sample Library, Cary, NC: SAS Institute Inc.
6. Brusilovskiy Pavel. M. and Hernandez Robert., “Rankings and Multi-Criteria Evaluation of Alternatives: Usage of Objective Data, Subjective Measurements and Expert Judgment” NESUG ‘97 Conference - Baltimore, MD (1997).

**Author Contacts**

Pavel Brusilovskiy  
 Conrail Corporation  
 2001 Market Street 6-D  
 P.O. Box 41406  
 Philadelphia, PA 19101-1406  
 (215) 209-7304  
 pbrusilovs@aol.com

Robert Hernandez  
 Conrail Corporation  
 2001 Market Street 14-C  
 P.O. Box 41403  
 Philadelphia, PA 19101-1403  
 (215) 209-4274  
 RHer1@aol.com