STOCHASTIC PROGRAMMING FOR TRAIN DISTRIBUTION IN A METRO TRANSPORTATION SYSTEM

Jorge A. Garcia[(a)], K Ponnambalam[(b)]

[(a),(b)] University of Waterloo, Ontario, Canada

[(a)] ja4garci@uwaterloo.ca, [(b)] ponnu@uwaterloo.ca

ABSTRACT

We present a mixed integer linear programming formulation with an application for the optimal train distribution problem. The objective was to assign trains to metro routes so that the sum of deficit capacity costs (i.e., passengers that cannot be carried) and the overcapacity costs (i.e., the number of empty seats) in a metro transportation system were minimized. We selected Mexico City's Metro Transportation System (MTS) as a study case. We narrowed our model to the operation and demand of peak hours. First, we solve a deterministic model using mean values for the parameters. Then, we built a two-stage stochastic model to include uncertainty into the parameters and we calculated estimates for mean and variance using maximum likelihood estimation. We discuss the results and compare the solutions for the four cases that we created. Finally, we propose an extension of the model that includes a time index.

Keywords: stochastic programming, train distribution, transportation system, integer programming

1. LITERATURE REVIEW

Train distribution is related to frequency optimization since there is a direct connection between number of trains and frequencies of rides. The use of linear models for frequency optimization in transportation has been widely studied; Yoo (2010) proposed an approach for a multi-modal transportation in which the problem is divided into two sub-problems, in the first they solve the route-selection problem from the passenger perspective, and for the second they find the optimal frequency from the operator perspective. Zhou et al. (2005) proposed a bi-level formulation in which the expected profit is maximized. Fernandez et al. (2008) proposed a methodology to deal with public transportation design and showed an application for the city of Santiago, Chile. Something that has been disregarded is the trade-off associated with non-uniform demand along the route, i.e., to have a higher demand in some particular stations in one direction than the rest of the line. We believe that measuring the overcapacity under such conditions would be useful for policy decision-making.

2. INTRODUCTION

Collective transport systems are a better mobility solution for medium to large distances than private cars within highly populated metropolis. Mexico City has more than 10 million people that are moving every day within the city. Its Metro Transport System (MTS) moves more than five million passengers every day. A key factor in the operation is the correct distribution of trains to meet the demand at every station; however, there is a fixed capacity in the whole system. So, the problem is how to assign trains to subsystems (i.e., routes) so that the operation is optimized.

We would like to present the following example to show the approach we used for the train distribution problem. Presented here is a five-station metro route with uneven flows between stations/directions in the studied time period.

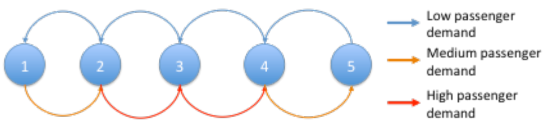

Figure 1. Example of a five-station metro route

Demand from station one to five is higher than that in the opposite direction (i.e., from five to one), also, stations two and three have a higher demand. This situation is common in transportation systems: to have a high demand of service in one direction in a certain period while the opposite direction has a low demand. One question that arises from the example: *is it worthy to have a high and fixed capacity in the whole route just to meet the demand of two stations?* Although it is true that the passenger flow changes over the day, the traffic demand is always concentrated in a subset of stations. As described above, we find compelling the formulation of a model that not only penalizes the unsatisfied demand at some station, but also the capacity not used on the rest. We believe this could be an interesting approach to evaluate the tradeoffs of having a fixed capacity on both sides of the route. All the routes we considered (routes 1 to 9) have trains with the same characteristics (e.g. pneumatic wheels, size, etc.).

3. METHODOLOGY

We followed the next methodology:

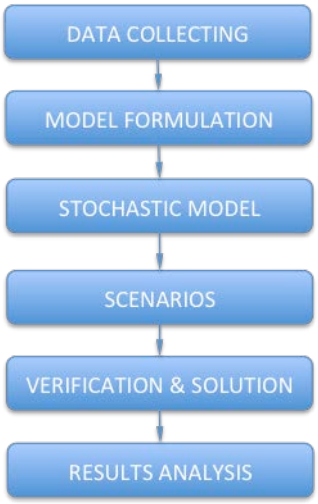

Figure 2. Methodology to address the optimal distribution problem

3.1 Data Collecting

Inflow passenger data per station and the number of trains were obtained from the Mexico's MTS site, which has data from 2012 to date, and it is available for public consultation.

3.2 Model Formulation

The process of model formulation started from the creation of one simple model and gradually we added more variables. First we present the deterministic model: objective function, decision variables, etc.

3.3 Stochastic model

We show the parameters that have uncertainty and the procedure we followed to handle their randomness Passenger inflow per station and service capacity were the main two parameters we estimated. Below we comment about the process and assumptions made.

3.4 Scenarios

For the stochastic model, we tried different number of scenarios. Using Monte Carlo sampling method we produced realization of the demand series. Since we are interested in the trade-offs between satisfying demand and overcapacity service we created four cases: the first two consist in finding the optimal distribution without penalizing the exceeding capacity; the other two, do penalize the exceeding capacity. The scenarios are explained in detail below.

3.5 Verification & Solution

In this step we first selected the software to write and solve the model. We chose Matlab due to its easy code writing and data retrieval utilities.

3.6 Result analysis

We discuss and compare the optimal solutions for both of the models, the deterministic and the stochastic.

4. MODEL FORMULATION

4.1 Decision variables

x_{ijk} is the number of passengers being transported from station *i* to station *j* on route *k* in the given time period. It includes the passengers from the previous station *(i-1)* plus the current passengers at station *i*.

v_k is the number of trains assigned to *k* route. It is the main decision variable since it determines the distribution of trains over the routes. It is an integer variable.

4.2 Auxiliary variables

d_{ijk} is the number of passengers not served at station *i* in direction to station *j* on route *k* in the given time period. It is used for quantifying the deficit of the service.

e_{ijk} is the number of empty seats at station *i* to station *j* on route *k* in the given time period. It is an auxiliary variable used to quantify the exceeding capacity.

4.3 Objective function

We used a linear cost function

$$\text{Min } z = \pi_1 \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{m} d_{ijk} \right) + \pi_2 \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{m} e_{ijk} \right) \qquad (1)$$

The objective function minimizes the total sum of the deficit capacity cost (π_1) and the exceeding capacity cost (π_2) occurred at every station and route of the metro system. By penalizing both the deficit (d_{ijk}) and the overcapacity (e_{ijk}) we are forcing the model to find an optimal tradeoff between exceeding the demand and exceeding the capacity. We set π_1=1 and π_2=0 for A and B scenarios, and π_1=6 and π_2=1 for III and IV scenarios, respectively.

4.4 Restrictions

The first set of restrictions is in the form

$$b_{ijk} + p_{ij} * x_{i-1ik} = x_{ijk} + d_{ijk}, \quad \forall i = 1,2,...,n-1, \quad \forall k \qquad (2)$$

The equation states that the quantity of passengers arriving at station *i* (represented by b_{ijk}) plus a percentage p_{ij} of passengers travelling from the previous station (*i-1*) is equal to the number of passengers being transported from *i* to *j* plus the deficit variable d_{ijk}.

$$x_{ijk} \leq C_k * v_k, \quad \forall i, j, k \qquad (3)$$

Equation (3) ensures that the rate of passengers travelling from *i* to *j* at *k* route is less or equal to the capacity of the train (C_k) times the number of trains (v_k). We show how C_k is obtained below.

$$C_k * v_k - x_{ijk} = e_{ijk}, \quad \forall i, j, k \qquad (4)$$

Equation (4) assigns the exceeding capacity to the overcapacity variable (e_{ijk}). From the restriction (3) we know that x_{ijk} is less or equal to C_k*v_k, so e_{ijk} will be zero or positive.

$$\sum_{k=1}^{m} v_k = T \tag{5}$$

Restriction (5) states that the sum of all assigned trains must be equal the total number of trains available (T).

4.5 Deterministic model

The complete deterministic model is

$$\text{Min } z = \pi_1\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} d_{ijk}\right) + \pi_2\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} e_{ijk}\right)$$

$$b_{ijk} + p_{ij} * x_{i-1ik} = x_{ijk} + d_{ijk} \qquad \forall i = 1,2,\ldots,n-1, \quad \forall k$$

$$x_{ijk} \le C_k * v_k \qquad \forall i,j,k$$

$$C_k * v_k - x_{ijk} = e_{ijk} \qquad \forall i,j,k$$

$$\sum_{k=1} v_k = T$$

$$x_{ijk} > 0 \qquad \forall i,j,k$$

$$v_k \in Z^+$$

The deterministic model has a total of 819 variables and 811 constraints.

5. STOCHASTIC MODEL

The deterministic model does not consider that demand and capacity are variable. Demand in one day may be different from that on another day. The capacity of the route depends on the number of train loops, which are not always the same due to technical reasons or delays. Stochastic models introduce randomness into parameters. Below we discuss how uncertainty was introduced into the model's parameters.

5.1 Demand (b_{ijk})

This parameter represents the number of passengers at station *i* moving to station *j* along *k* route. The first step was to calculate the mean and standard deviation.

From the historical data, we did an exploratory graphical analysis. We discovered a positive trend in the demand series (Figure 3). We removed the trend by differentiating one time (Figure 4).

The histogram of the differentiated series resembles a Normal distribution (Figure 5), so we decided to assume the demand behaves in that manner. We calculated an estimate for the mean and variance using Maximum Likelihood estimation (See Appendix B) for every route.

Then, we studied the correlation in the series using the sample autocorrelation function and the sample partial autocorrelation function. We noticed that the series cuts off after lag 1 in the autocorrelation function, implying that there is a negative correlation only between time *t* and *t+1* (Figure 6). The partial autocorrelation function shows that the process is dying off, which confirms a negative correlation (Figure 7).

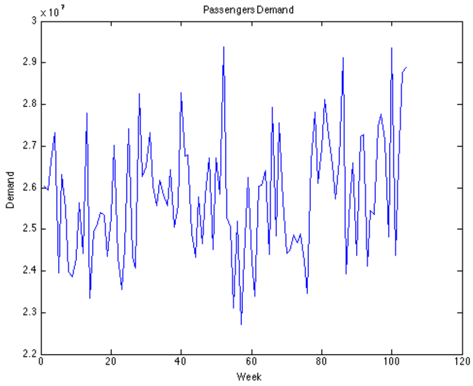

Figure 3. Weekly passengers demand from January 2012 to December 2013

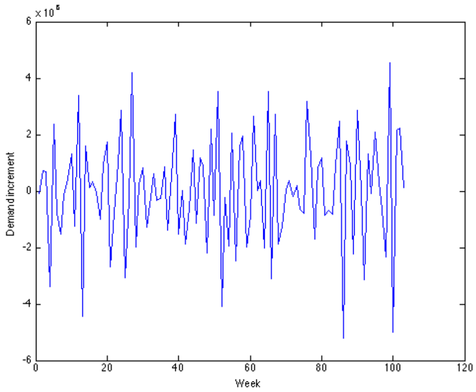

Figure 4. Differentiated demand series

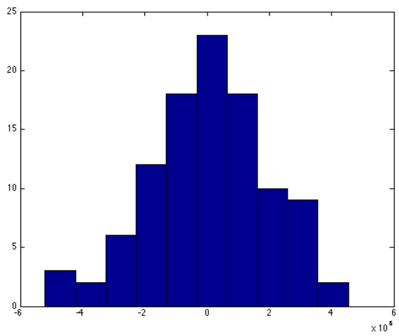

Figure 5. Histogram of the demand after differentiating

Now, we have data about the number of passengers entering at station *i* in route *k*, but we do not know how many of them are moving in which direction. We determined these two numbers by using the linear combination:

$$b_{ik} = \lambda b_{ijk} + (1-\lambda) b_{hik}, \qquad 0 \le \lambda \le 1$$

b_{ik} is the known number of passenger inflow in the station *i*, route *k*; b_{ijk} is the portion of b_{ik} moving in the direction of *j*; b_{hi}, is that moving in the direction of *h*. We assume lambda decreases linearly from 1 in station *i=1*, to 0 in station *i=n* (terminal).

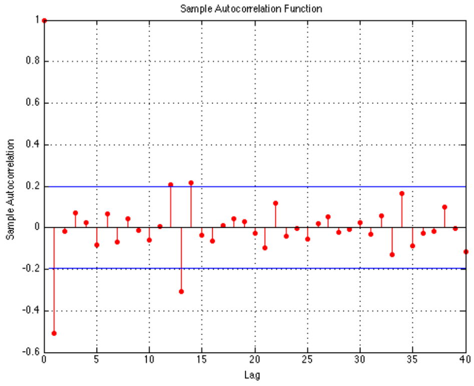

Figure 6. Sample auto correlation function

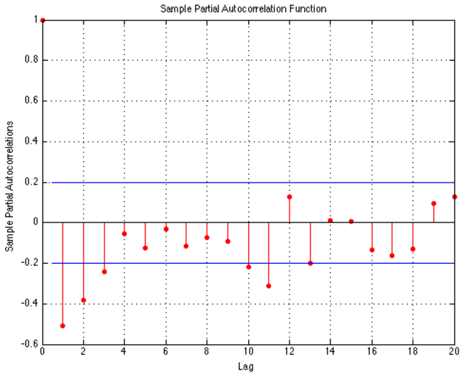

Figure 7. Sample partial autocorrelation function

5.2 Changes in the objective function

Following the formulation of the two-stage stochastic programming, we added two more variables δ_{ijk}^{s} and ε_{ijk}^{s} to absorb the deficit or surplus of demand for every scenario *s*. We penalized those values with the same weight we gave to the deterministic variables for deficit (π_1) and surplus (π_2) times 1/S; S is the total number of scenarios:

$$\frac{\pi_1}{S}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{s=1}^{S}\delta_{ijk}^{s}\right)+\frac{\pi_2}{S}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{s=1}^{S}\varepsilon_{ijk}^{s}\right)$$

Since it is a minimization model, these costs have a positive sign on the objective function.

5.3 Changes in restrictions

The only set of restriction that changed is the first one. We added the two-stage variables of deficit (δ_{ijk}^{s}) and surplus (ε_{ijk}^{s}):

$$b_{ijk}+p_{ij}*x_{i-1ik}+\varepsilon_{ijk}^{s}-\delta_{ijk}^{s}=x_{ijk}+d_{ijk}$$
$$\forall i=1,2,\ldots,n-1,\quad \forall k,s$$

5.4 Service Capacity (C_k)

This parameter represents the capacity of service per unit of time and per train. In this work we considered the unit of time as one hour, so for every station we calculated C_k by a function of the length and the number of stations. A comparison was made between the results obtained using this formula and those provided by the City Metro – Map & Route Planner App (See Appendix A).

The time of one circuit loop for every train is

$$t=2(\alpha*s+\beta*l)$$

where *s* is the number of stations and *l* is the total length of the route in meters. α and β are random variables for waiting time and speed, respectively. We assume they have a uniform distribution over the interval [0.8μ, 1.2μ] where μ is the mean value. We also assume that the mean value for α is 1.5 minutes, and that for β is 35.5 km/hour. So, we calculate the expected time for every route and then, use that value to obtain the expected capacity service:

$$E[t_k]=3s_k+\frac{6}{1775}l_k \qquad (6)$$

So, equation (6) is the formula to calculate the expected time of one circuit for every route k.

Now we proceeded to calculate C_k and do the dimensional analysis:

$$\frac{60[\min/hour]}{t_k[\min]}*tc\left[\frac{passengers}{train}\right]=\left[\frac{passenger}{hour-train}\right]=C_k$$

Finally, we have

$$C_k=\frac{35{,}500\,tc_k}{1775s_k+2l_k} \qquad (7)$$

The capacity (tc_k) is 1,020 passengers per train in routes four and six, and 1,530 for the rest of the routes. The results are shown in Table 1.

We used the $C_k^{0.98}$ value for restrictions 2 and 3, to ensure the solution will satisfy the capacity requirement 98 out 100 cases.

Table 1. Summary data of the C_k (capacity)

Route	Stations s	l (m)	t_k	C_k (mean)	$C_k^{0.98}$ (98% conf.)
1	20	16,804	116.8	786	646
2	24	20,863	142.5	644	529
3	21	21,428	135.4	678	556
4	10	9,513	62.1	984	809
5	13	14,435	87.8	1,045	857
6	11	11,434	71.6	854	701
7	14	17,011	99.5	923	756
8	19	17,679	116.7	786	646
9	12	13,033	80	1,147	940

5.3 Two-Stage Stochastic Model

The complete two-stage stochastic model is

$$\text{Min } z = \pi_1\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} d_{ijk}\right) + \pi_2\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} e_{ijk}\right)$$

$$+\frac{\pi_1}{S}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{s=1}^{S}\delta_{ijk}^{s}\right)+\frac{\pi_2}{S}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{s=1}^{S}\varepsilon_{ijk}^{s}\right)$$

$$b_{ijk}+p_{ij}*x_{i-1ik}+\varepsilon_{ijk}^{s}-\delta_{ijk}^{s}=x_{ijk}+d_{ijk} \quad \forall i=1,\ldots,n-1, \quad \forall k,s$$

$$x_{ijk}\le C_k*v_k \qquad \forall i,j,k$$

$$C_k*v_k-x_{ijk}=e_{ijk} \qquad \forall i,j,k$$

$$\sum_{k=1} v_k = T$$

$$x_{ijk}\ge 0 \qquad \forall i,j,k$$

$$\delta_{ijk}^{s},\varepsilon_{ijk}^{s}\ge 0 \qquad \forall i,j,k$$

$$v_k\in Z^{+}$$

The number of variables and restrictions depends on the number of scenarios; for the largest case we tried (35 scenarios), it has a total of 19,719 variables and 9,991 restrictions.

6. SCENARIOS

We used Monte Carlo sampling method to produce *s* scenarios (realizations), each of them with probability of 1/S. We assumed the demand has a normal distribution; the mean and the variance were obtained from the historical data (see Appendix B). We solved the model for 2, 5, 10, 20, 25, 30, and 35 scenarios. From the tenth onwards, the optimal solution did not change.

In order to compare the solutions of the deterministic and stochastic models, we produced four cases:

1. Deterministic model; assigning π_1=1, π_2=0, and average values in demand and service capacity.
2. Stochastic model; assigning π_1=1, π_2=0, and using 35 scenarios for the demand and 98% of confidence in service capacity.
3. Deterministic model; assigning π_1=6, π_2=1, and average values in demand and service capacity.
4. Stochastic model; assigning π_1=6, π_2=1, and using 35 scenarios for the demand and 98% of confidence in service capacity.

In III and IV cases we are establishing that one occupied seat is six times more important than one available space (i.e., π_1=6 and π_2=1).

7. RESULTS

The results for the optimal train distribution are shown in Table 2. To make comparable the solutions, we associated a cost for the unmeet demand (Table 3). The total cost is equal to the sum of the costs: *c*(capacity-demand)*, for every route, we set *c=1*.

Table 2. Distribution of trains for every route

Route	Current	Optimal solution (trains)			
		I	II	III	IV
1	30	29	29	35	24
2	24	37	42	38	40
3	29	33	33	30	28
4	8	4	3	4	9
5	15	12	7	9	10
6	9	10	5	9	13
7	19	12	12	15	15
8	18	18	18	15	15
9	17	14	20	14	15

Table 3. Passengers capacity per hour

Route	Optimal solution (passengers)			
	I	II	III	IV
1	22,794	22,794	27,510	18,864
2	23,828	27,048	24,472	25,760
3	22,374	22,374	20,340	18,984
4	3,936	2,952	3,936	8,856
5	12,540	7,315	9,405	10,450
6	8,540	4,270	7,686	11,102
7	11,076	11,076	13,845	13,845
8	14,148	14,148	11,790	11,790
9	16,058	22,940	16,058	17,205
Cost	65,725	66,102	65,977	64,163

We can see that the minimal cost occurs on the case IV, i.e., solving the two-stage stochastic model for 35 scenarios.

The mayor changes are in the route 2 and seven. All the cases have at least 13 more trains in the former and at least four less in the latter. The current train distribution seems to be uneven in regards to meeting the demand.

8. CONCLUSIONS

In this paper we proposed a two-stage stochastic mixed integer program to assign train to routes in a

transportation system, minimizing a cost function, which penalizes both, deficit and exceeding capacity.
We found the function to be a compelling idea to assess the trade-offs associated to a non-uniform distributed demand in the system.

9. EXTENSIONS OF THE MODEL

One problem with the proposed model is that it is based on hour-average operation parameters. If we add a time index to *x*, *d* and *e* variables, then we can change the size of the time interval and make it arbitrarily small to obtain a more accurate result; however, the number of variables increases notably. So, we have for the stochastic model:

$$\text{Min } z = \pi_1\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{t=1}^{T} d_{ijkt}\right) + \pi_2\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{t=1}^{T} e_{ijkt}\right)$$

$$+\frac{\pi_1}{S}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{s=1}^{S}\sum_{t=1}^{T} \delta^s_{ijkt}\right) + \frac{\pi_2}{S}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{s=1}^{S}\sum_{t=1}^{T} \varepsilon^s_{ijkt}\right)$$

$$b_{ijkt} + p_{ijt} * x_{i-1jkt} + \varepsilon^s_{ijkt} - \delta^s_{ijkt} = x_{ijkt} + d_{ijkt} \qquad \forall i = 1,\dots,n-1, \quad \forall k,s,t$$

$$x_{ijkt} \le C_k * v_k \qquad \forall i,j,k,t$$

$$C_k * v_k - x_{ijkt} = e_{ijkt} \qquad \forall i,j,k,t$$

$$\sum_{k=1} v_k = Total$$

$$x_{ijkt} \ge 0 \qquad \forall i,j,k,t$$

$$\delta^s_{ijkt}, \varepsilon^s_{ijkt} \ge 0 \qquad \forall i,j,k,t$$

$$v_k \in Z^+$$

ACKNOWLEDGMENTS

We would like to thank the economical support from CONACYT and the Science, Technology and Innovation Office (SECITI) of Mexico City to make this work possible.

APPENDIX A. TRAIN LOOP TIMES

In Figure 8 we show a comparison between our formula t_k and the times calculated by a commercial application.

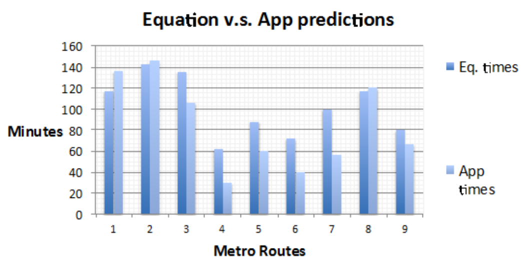

Figure 8. Comparison of one-loop times

The highest relative difference is on route 4 where the formula estimation is 107% higher than the App's. We believe that App predictions underestimate loop times in low-demand routes (4 to 7) possibly because it uses passenger demand to calculate the times and delays disregarding distance.

APPENDIX B. MONTHLY AVERAGE DEMAND

The estimates for mean and standard deviation were calculated from the historical data using the maximum likelihood estimation. The monthly results are shown in the following table:

Table 4. Monthly mean and standard deviation

Route	Mean	SD	CV
1	21,594,861	1,044,094	4.8%
2	24,901,980	851,348	3.4%
3	20,105,945	588,261	2.9%
4	2,413,326	166,630	6.9%
5	6,611,076	376,188	5.7%
6	4,184,297	481,499	11.5%
7	7,610,587	596,027	7.8%
8	11,471,053	397,340	3.5%
9	9,657,134	586,501	6.1%

In the box plot of the differentiating total demand series, we can see that it is highly concentrated around the median. The interquartile range has length of 2.5 millions, which represents approximately 28% of the range. It has long extreme values though. The median is around zero. Two outliers are shown in red.

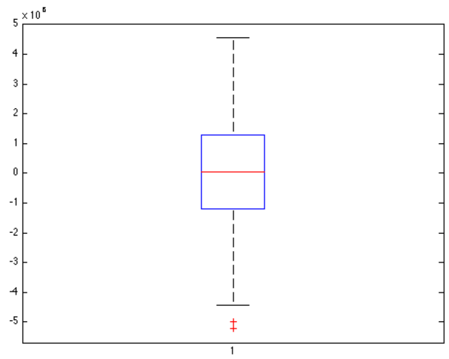

Figure 9. Boxplot of the observations after differentiating

REFERENCES

Fernandez E. et al. 2008, Demand Responsive Urban Public Transport System Design: Methodology and Application. Transportation Research Part A Vol. 42 No. 10: p 951-972

Yoo g., et al. 2010, Frequency Design in urban Transit Networks with Variable Demand: Model and Algorith. Transportation Engineering Vol. 14 No 3: p. 403-411

Zhou J., Lam, W., Heydecker B. 2005, The Generalized Nash Equilibrium Model for Oligopolistic Transit Market with Elastic Demand. Transportation Research B Vol. 39 No. 6: p. 519-544, 2005

AUTHORS BIOGRAFY

Jorge A. Garcia is a PhD student of Systems Design Engineering, University of Waterloo, working in the Design Optimization Under Uncertainty group.

K. Ponnambalam is a professor of Systems Design Engineering, University of Waterloo, and works in the areas of design optimization under uncertainty.