THE AMERICAN STATISTICIAN
2017, VOL. 71, NO. 3, 249–258: Teacher's Corner
https://doi.org/10.1080/00031305.2016.1256231

Taylor & Francis
Taylor & Francis Group

Resolving a Multi-Million Dollar Contract Dispute With a Latin Square

William B. Fairley[a], Peter J. Kempthorne[b], Julie Novak[c], Scott McGarvie[d], Steve Crunk[e], Bee Leng Lee[e], and Alan J. Salzberg[f]

[a]Analysis & Inference, Inc., Springfield, PA; [b]Mathematics Department, Massachusetts Institute of Technology, Cambridge, MA; [c]IBM Thomas J. Watson Research Center, Yorktown Heights, NY; [d]Bank of England, London, UK; University of Exeter, Exeter, UK; [e]Department of Mathematics and Statistics, San Jose State University, San Jose, CA; [f]Salt Hill Statistical Consulting, New York, NY

ABSTRACT

The City of New York negotiated a dispute over the performance of new garbage trucks purchased from a vehicle manufacturer. The dispute concerned the fulfillment of a specification in the purchase contract that the trucks load a minimum full-load of 12.5 tons of household refuse. On behalf of the City, but in cooperation with the manufacturer, the City's Department of Sanitation and consulting statisticians tested fulfillment of the contract specification, employing a Latin Square design for routing trucks. We present the classical analysis using a linear model and analysis of variance. We also show how fixed, mixed, and random effect models are useful in analyzing the results of the test. Finally, we take a Bayesian perspective to demonstrate how the information from the data overcomes the difference between the prior densities of the city and the manufacturer for the load capacities of the trucks to result in much closer posterior densities. This procedure might prove useful in similar negotiations. Supplementary material including the data and R code for computations in the article are available online.

ARTICLE HISTORY
Received March 2015
Revised September 2016

KEYWORDS
Analysis of variance; Bayesian; Fixed effect; Mixed effect; Random effect

1. Introduction

A statistically designed field experiment was developed to aid in negotiations between New York City's Department of Sanitation and the Mack Truck Company, a maker of waste collection vehicles (i.e., garbage trucks). A multi-million dollar contract for new vehicles between the city and the manufacturer called for the new trucks to load a minimum full-load of 12.5 tons of household refuse. The city doubted the capability of the trucks to load the required amount and withheld payment for the trucks. The New York City Law Department engaged the services of statisticians—two of the authors of this article—to help determine if the manufacturer had fulfilled the load specification.

2. The Dispute and Negotiations

2.1. Test of an Engineering Fix

When refuse is dumped into the opening at the rear of a truck, a hydraulic system drives a panel forward to squeeze it in. The city had relayed its concern over inadequate tonnage, causing the manufacturer to undertake an engineering fix to the hydraulics to increase tonnage. The manufacturer took the new design back to the city and offered to retrofit the trucks already delivered. But how good was the fix? The manufacturer took the position that the fix was adequate. The city retained a mechanical engineering consultant to study the proposed fix. Based on the results of his study, the city told the manufacturer that it was uncertain whether the fix would bring the trucks up to the load specification of the contract under actual operating conditions. The city then hired statisticians to develop and implement a test to determine whether the prospective fix was adequate.

The manufacturer wanted assurances that the trucks would be tested as soon as possible, trucks would be fully loaded when they weighed in, standard procedures would be followed, and trucks would be in good repair. The manufacturer agreed to the test with the provisos that they review the operational protocol, be permitted to observe, and receive the data and analysis at the same time as the city. The city agreed.

2.2. The Contract and Key Questions

Did the city interpret the contract to mean that every truck would always hold at least 12.5 tons when fully loaded? Could all but some fraction of the loads meet the specification? Further, environmental conditions affected the loads. Some routes had characteristically different densities of refuse. Different crews might have loaded more than others. The dispute might have been avoided had the contract answered such questions with unambiguous acceptance criteria (see Juran and Godfrey 1988, sec. 46, "Acceptance Sampling"). The test was designed to give useful information to the parties about the capacities of the trucks apart from the contribution of environmental factors in the loads. The three key questions the test was designed to answer were:

CONTACT Julie Novak jenovak@us.ibm.com IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

- Do trucks on average load 12.5 tons or more?
- Are loads under 12.5 tons due to differences between trucks or to differences in the environments in which trucks are run?

A somewhat less important question was whether the truck capacities differ.

3. The Test

3.1. *Controlling for Other Factors in the Test*

The city believed that differences in performance might be due to: (1) different routes and (2) different days of the week. Routes vary in frequency of pickup and density of the resident population. Compactability of waste may vary between days—particularly between the beginning and end of the week, since there is no regular weekend pickup and weekend refuse differs from weekday refuse. The test would be biased if the truck-runs were concentrated on unusual routes or unusual days. The design of the test avoided such bias by running trucks in broadly different classes of routes and on each weekday. An alternative test would be to load up trucks at a dump and check the tonnage. However, in this test one would have to decide the composition of the load because loads differ. A load of inner spring mattresses and a load of discarded barbells are going to have different densities. Also, in the eyes of the city, if crews in actual operation did not get 12.5 tons, the trucks would be deficient.

The city divided routes into classes that differed in frequency of pickups (2, 3, or 5 times per week) and in the density of the resident population (low, medium, or high). For truck runs, the city selected a route within each of the five classes, and conducted the test on each of the five chosen routes on all five weekdays. Table 1 summarizes the route information.

The protocol for the test should ensure that truck loads were not influenced by operating procedures that were atypical. The city and the manufacturer wanted to find out whether the trucks themselves measured up, not whether the crews measured up. A protocol that the truck runs should follow was written to set out daily procedures expected of a diligent crew. The protocol provided for the presence on each run of a foreman and mechanic, and the recording of information about the run on a report form, preinspection of trucks prior to a run, and procedures for accurately determining the load. Importantly, every truck was driven on a run until it was full, so that smaller loads would not reflect simply an inadequate filling. The full-load weight was determined by weighing a truck before and after dumping its full load, and taking the difference. Calibration of the scales was required prior to use.

Table 1. Routes.

Class	Route	Frequency of pickup	Population density
Queens West	1	2x	Low (L)
Brooklyn West	2	2x	Medium (M)
Brooklyn North	3	3x	Medium (M)
Manhattan West	4	3x	High (H)
Manhattan East	5	5x	High (H)

Table 2. The Latin Square of truck assignments for the 5 day period under examination. This same permutation has been applied to both sets of trucks; A1, B1, . . ., E1 and A2, B2, . . ., E2.

	Day				
Route	M	T	W	TH	F
1 Queens West	C	A	B	E	D
2 Brooklyn West	E	D	C	A	B
3 Brooklyn North	A	E	D	B	C
4 Manhattan West	D	B	A	C	E
5 Manhattan East	B	C	E	D	A

3.2. *The Latin Square Design for Truck Runs*

If a single truck were run on every combination of five routes and five days, the total number of runs required in the test of that truck would be 5 × 5 or 25. Since 10 trucks were used, that is, 10 × 25, or 250 truck runs. This would be expensive, and a logistical nightmare. Perhaps it would be downright impossible, because on any given day a truck would have to be moved all over the city to reach five different routes, weighed, and moved into place for the following day's test. Fortunately, the Latin Square method of assigning truck-runs to the factors of route and day cuts down the number of runs required by a factor of 5, so that instead of 250, only 50 are needed. The efficiency of the Latin Square design made the test possible.

Table 2 displays the Latin Square used in the actual test. Five trucks are designated A, B, C, D, and E. Each letter represents a truck (a truck level). For example, C in the top left of the array of letters means that truck C operated in Route 1 (Queens West) on Monday. As noted below there were actually a total of 10 trucks—two for each of the five letters. Table 3 shows how the 10 trucks are labeled.

Two trucks were assigned at random to each of the treatment categories in the Latin Square. All 50 truck-runs were carried out in the same week. The pattern of two repeated Latin Squares (with one of the treatment or blocking factors—trucks— having different levels in the two squares) is sometimes termed a Latin Rectangle (see Casella 2008, p. 121). It is important to note that the repeated assignment used here differs from what one would apply in classical experimental design. Typically, one arrangement of the Latin square would be used for trucks 1 to 5 and then a distinct permutation would be applied for trucks 6 to 10. Pairs of trucks would therefore not stay together throughout all five route-day combination. In our context, this was not possible, due to the way in which the city conducted the experiment.

Table 3. A table summarizing the tons of refuse recorded from the test. Two groups of five trucks were assigned to runs according to the same Latin Square. This resulted in a total of 50 runs, made by 10 different trucks.

	First set of trucks (A1, B1, C1, D1, E1) Day					Second set of trucks (A2, B2, C2, D2, E2) Day				
Route	M	T	W	TH	F	M	T	W	TH	F
1	14.0	15.3	13.5	13.1	13.3	13.4	15.2	12.6	12.7	11.7
2	14.8	15.1	13.7	14.5	13.0	14.7	15.6	13.9	14.1	13.8
3	14.5	15.2	13.8	13.5	12.1	13.8	14.5	14.2	13.0	12.4
4	14.2	14.7	13.3	12.5	12.7	13.4	13.5	13.3	12.0	13.4
5	13.2	14.4	15.6	14.0	12.4	15.2	15.6	13.3	12.4	13.0

The Latin Square has the property that each level of the treatment factor (trucks) is run once—and only once—at each level of the Route factor and at each level of the Day factor. The important consequence here of using the Latin Square to assign truck-runs is that for each truck the average of its loads is not biased by being run on a skewed selection of routes or days.

The Latin Square design was also thought to work well here because the city's engineering consultants believed that the trucks with the hydraulic fix, if they worked well, would work about the same no matter where in the city or with what type of refuse they were used.

One important point to note throughout the analysis that follows, is that we have been required to assume that interactions between factors are absent. Given the experimental restrictions under which this test had to be conducted there is no way of verifying whether these interactions actually exist—the data available in the Latin square design are simply too small to check this condition. One should therefore be mindful of this assumption when interpreting the results, however we believe that this is a reasonable approximation to make, given our understanding of the problem. A physical mechanism for generating interactions between day, route, and trucks is possible, however we believe these effects will be negligible.

The use of a Latin Square design here is unusual in that, instead of the goal being to see if one or more of the treatments (the trucks) are better, the goal is to see how similar they are. That is, the trucks are expected to load 12.5 tons or more and, in addition, each is expected to load an amount similar to the others. This is not a problem, as the same machinery that can discover how different the treatments are can also discover if they *are* different.

3.3. Data from the Test

Table 3 gives the results of the test—the tons of household refuse loaded on each truck-run. The values in the table were obtained from reports completed by supervisors and mechanics each day. Figure 1 displays a histogram of the data given in Table 3.

On average, the 50 loads in the test easily met the contract specification of 12.5 tons. However six runs, or 12% of the 50 loads are below the 12.5 ton specification. If we look at the six truck loads out of the 50 that are below the specification, two are on Thursday and four are on Friday. These two days have the lowest average per-day tons. It is possible that these six loads are not below the specification because the trucks on average could not load 12.5 tons, but rather, at least in part, because they are run on low-average-density days.

The original dataset contained four missing values of tons loaded, due to equipment breakdowns. The missing values were at the following route/day/truck factor level combinations: 1/Tu/A1; 1/Tu/A2; 4/W/A1; and 4/M/D1. To allow one to carry out a standard analysis procedure, which requires a complete dataset, we have chosen to replace these missing values with a single imputation using a regression model. Since the cause of their absence is a mechanical error, we believe that the points are missing at random, and we sample candidate values as if they were from the same distribution as the observed points.

A linear regression is fitted to the observed tonnage data using day, route, and truck as factors. We then sample the missing values from a normal distribution centered at their respective fitted means and having standard deviation equal to the standard error obtained from the regression model. It is important to emphasize that one should not just fill the missing observations with the average values of the observed data as this

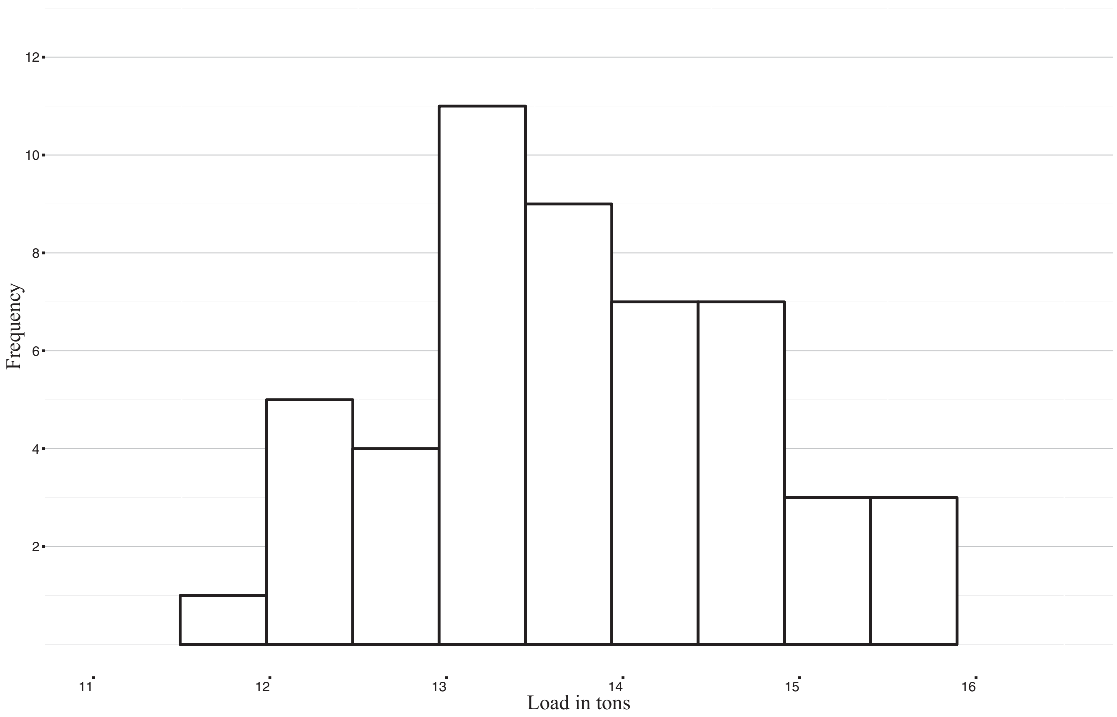

Figure 1. Histogram of the 50 loads given in Table 3.

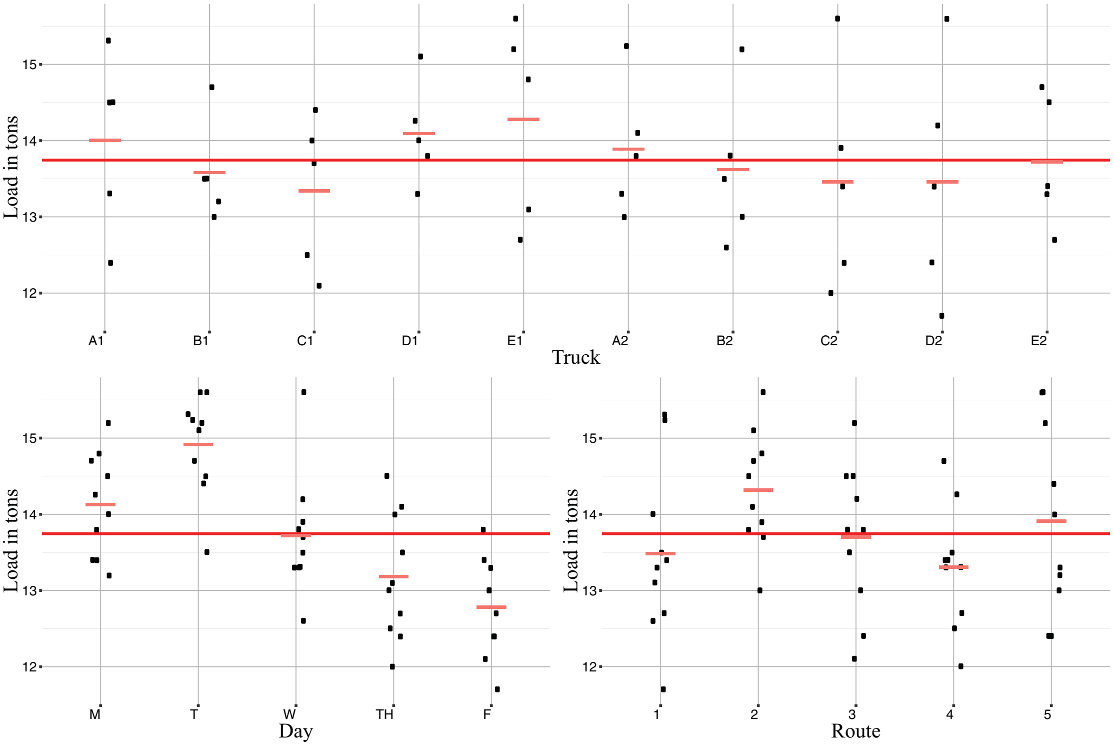

Figure 2. Loads by route, day, and truck showing the population mean above 12.5 tons, and modest differences between trucks and between routes, but larger differences between days. The horizontal line across the entire plots show the grand mean of the loads, 13.74 tons. The small horizontal lines show the means for the levels for each specific factor.

method would result in the data having artificially small observational error for the imputed values.

For pedagogical reasons we chose to limit the imputation procedure to the relatively simple method described above. A fuller treatment of the missing data points would be to use a multiple imputation procedure in which one samples from the residual distribution, estimates the observed effects from the model, and then repeats this process many times. One can then examine the distributions of fitted effects and determine how these vary under different realizations of the data.

4. Fixed, Mixed, and Random Effects Models of Loads

We can take an exploratory point of view to describe the data by their grand mean and by effects due to the factors route, day, and truck—seeing these quantities as successive layers of fit (see Hoaglin, Mosteller, and Tukey 1991). Figure 2 displays the data obtained by route, day, and truck. The solid horizontal line in each panel denotes the grand mean, while the small horizontal lines indicate the means for the specific factors. The effects can be considered the deviations of the means for the specific factors from the grand mean of all the loads.

4.1. *Average Load and Truck Capacities Under the Fixed Effects Model*

Letting y_{rdt} be the load in tons of a truck run on route r, day d, and with truck t, we can describe the data as a sum of their grand mean plus each of the effects of route, day, and truck as follows:

$$y_{rdt} = \mu + \gamma_r + \tau_d + \rho_t + \varepsilon_{rdt}, \tag{1}$$

where $r = 1, 2, \ldots, 5$, $d = 1, 2, \ldots, 5$, and $t = 1, 2, \ldots, 10$; μ is taken to be the true mean load of all 50 truck runs; γ_r is the effect of the rth route, τ_d is the effect of the dth day, ρ_t is the effect of the tth truck. ε_{rdt} is an “error” term, which models unobserved factors as well as pure random variability. When the “model” is referenced, the context will govern whether Equation (1) or the mean model $(\mu + \gamma_r + \tau_d + \rho_t)$ is meant. Estimated effects from a classical analysis of variance (ANOVA) fit corresponding to the parameters of the model (1), in order of the addends, are, respectively:

$$\bar{y}_{...}, \quad \bar{y}_{r..} - \bar{y}_{...}, \quad \bar{y}_{.d.} - \bar{y}_{...}, \quad \bar{y}_{..t} - \bar{y}_{...} \tag{2}$$

with $r = 1, 2, \ldots, 5$, $d = 1, 2, \ldots, 5$, and $t = 1, 2, \ldots, 10$. It can be seen that the addition of the sample grand mean, $\bar{y}_{...}$, to the last three terms of (2) gives the sample means of the three factors, that is, $\bar{y}_{r..}$, $\bar{y}_{.d.}$, and $\bar{y}_{..t}$, which are estimates of $\mu + \gamma_r$, $\mu + \tau_d$, and $\mu + \rho_t$, respectively.

Figure 3 shows the errors resulting from the fit of the fixed effects model defined in Equation (1). Upon examination of the residuals there is no evidence of outliers or obvious patterns in the data and so on initial inspection, the model appears to describe the data adequately.

A major objective of the designed experiment was to determine if the average load would indeed be greater than the specification. The mean of the 50 loads is 13.74 tons, which is substantially above the specification of 12.5 tons. The standard deviation of the loads themselves is 1.02 tons, and the standard error of the mean load—under the model $y_i = \mu + \epsilon_i$—is 0.14 $(0.14 = 1.02/\sqrt{50})$. The mean load therefore is over eight standard errors above the specification. After the fact, the added precision of the mean afforded by the experimental design is

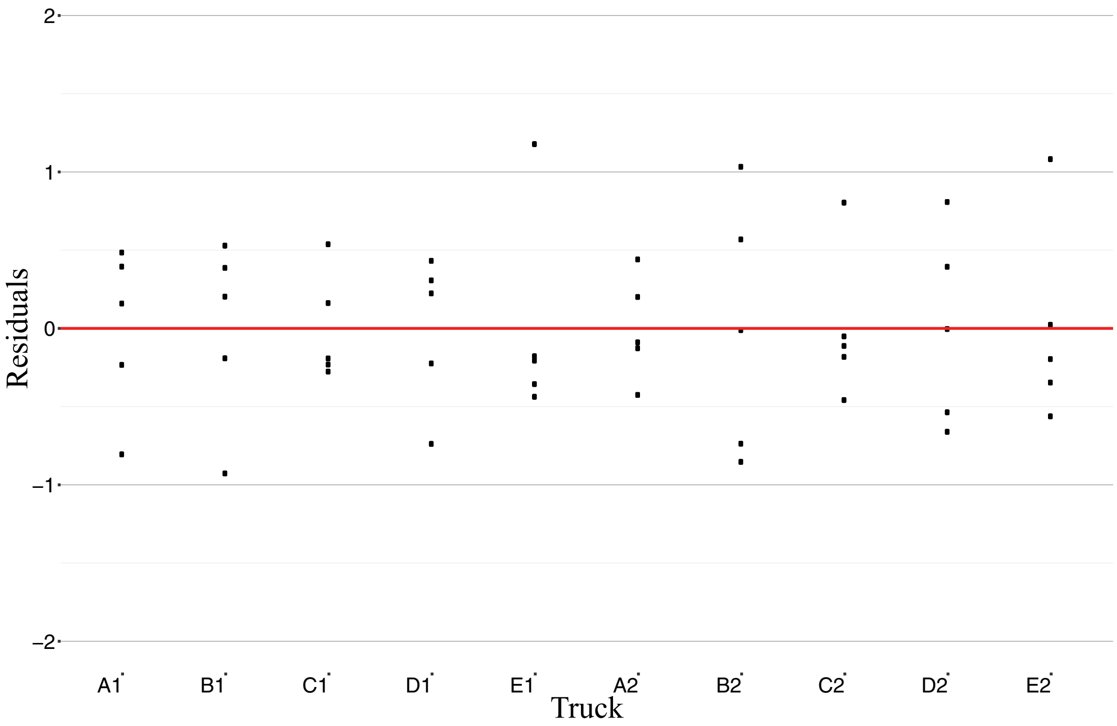

Figure 3. The residual errors resulting from the fitted fixed effects model defined in Equation (1).

not required to reliably determine that the mean load meets the contract specification.

Although the grand mean is reliably above specification, there could be some trucks whose mean loads are below specification. That is, inadequate truck capacities could be responsible for loads below the specification. The estimated means of the 10 trucks are 14.00, 13.58, 13.34, 14.09, 14.28, 13.89, 13.62, 13.46, 13.46, 13.72—having a range of 0.94 tons and all being above the contract specification. Figure 4 displays the estimated mean of each truck as indicated by the solid points along with two horizontal lines drawn at the specification of 12.5 tons and at the grand mean, 13.74 tons, of all 50 loads.

Confidence intervals for the truck means can be constructed under the assumption of a probability model for the loads in which the fixed effects model at (1) is augmented with the distributional assumptions that the errors are independent and normal:

$$\varepsilon_{rdt} \sim N(0, \sigma^2)\ \forall r, d, \text{ and } t. \tag{3}$$

Figure 4 also contains 50% and 90% confidence intervals for the true mean load, $\mu + \rho_t$, similar to the procedure outlined by Gelman (2005). The intervals are all above the target of 12.5 tons and so the proposition that the target is met on average by each of the trucks in the test is supported. Hence, the statistical treatment of this experiment, as outlined in this document, allows one to separate the sources of variation attributed to route, day, and truck factors and has made it possible to establish with high confidence, that the truck means individually meet the specification.

In support of the probability model of the loads just specified, there is no reason to suspect that the errors are not independent. A normal score, Q-Q plot of the residuals (not shown) indicates normality is a reasonable assumption, and Figure 3 does not suggest heteroscedasticity in the error terms.

The city would have some interest in knowing whether the truck capacities differed between trucks. The analysis of variance table (ANOVA) shown in Table 4 sets out the statistics through which this question can be addressed. Adopting the framework of hypothesis testing, the F-statistic for trucks tests the hypothesis that the truck effects are zero. It is not statistically significant (p-value 0.33182), so the null hypothesis that the truck effects are zero is not rejected.

For completeness, we include Table 5, which gives the least-square estimated fixed effects for the factors.

4.2. Average Load and Truck Capacities Under the Mixed Effects Model

Both parties were interested in the capacity of all 120 trucks that the manufacturer would be delivering, not just the 10 selected for the test. Further, they were interested in loads over all routes, not just over the 5 selected by the city from the 26 available. This motivates analysis of a model in which the effects of the 10 trucks are considered to come from a distribution of trucks, and the effects of the five routes are considered to come from a distribution of routes. We have observations for all five weekdays however and so we do not initially model these as originating from a distribution of days. As such the "mixed effects" model assigns distributions to trucks and routes, but not to days. Specifically, truck and route effects are specified as coming from a normal distribution both of which have a mean of zero and unknown variance given by

$$\begin{aligned} \gamma_r &\sim N\big(0, \sigma_R^2\big), \quad r = 1, 2, \ldots, 5, \\ \rho_t &\sim N\big(0, \sigma_T^2\big), \quad t = 1, 2, \ldots, 10. \end{aligned} \tag{4}$$

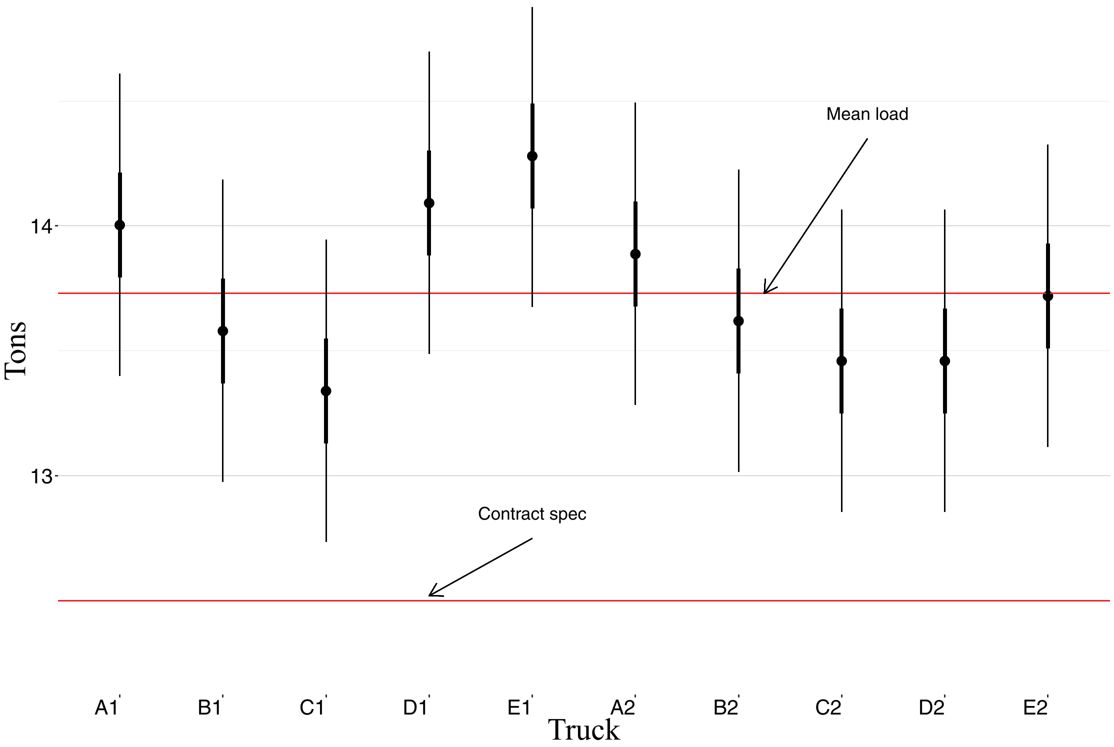

Figure 4. Observed truck means with 50% (thicker bars) and 90% (extending to the ends of the thinner bars) confidence intervals. The intervals have been derived from the residual variance assuming that we know μ and ρ_t. In other words, the variability in the intervals includes the variability from the route and the day. All 10 truck means and their intervals are above the specification of 12.5 tons. Since under the fixed effects model, the least-square estimate of $\mu + \rho_t$ is $\bar{y}_{..t}$, the standard error of $\bar{y}_{..t}$ is derived by substitution of the model at (1) into $\bar{y}_{..t}$, and taking the variance. The result is the value $\sqrt{\sigma^2/5}$. σ^2 is estimated by the mean square for error in the ANOVA, 0.4018 (see Table 4). That is, the standard error of $\bar{y}_{..t}$ is estimated to be $\sqrt{0.4018/5} = 0.2835$.

The parameters σ_R^2 and σ_T^2 take the place of the 5 route effects and the 10 truck effects in the fixed effects model, respectively. The complete model for the three factors under examination are therefore described by two variance components for route and truck effects, and five fixed effects for each day. Equations (1), (3), and (4) therefore define the mixed effects model of truck capacities. The parameters of the mixed effects model have been estimated by (restricted) maximum likelihood.

Table 6 displays the estimated variance parameters, σ_R^2 and σ_T^2, for the mixed effects model.

Table 4. ANOVA table of results.

Source	SS	df	MS	*F*	*p*-Value
Route	6.193	4	1.548	3.853	0.01149
Day	27.651	4	6.912	17.206	1.26E-07
Truck	4.321	9	0.480	1.195	0.331
Error	12.856	32	0.402		
Total	51.021	49			

Table 5. The estimated fixed effects for the factors. These are estimated by least squares from the fixed effects model in Equation (1) subject to the constraints that the sum of the effects for each factor is zero.

Route	1	2	3	4	5
Effect	−0.26	0.58	−0.04	−0.44	0.17
Day	M	T	W	TH	F
Effect	0.38	1.17	−0.02	−0.56	−0.96
Truck	A1	B1	C1	D1	E1
Effect	0.26	−0.16	−0.40	0.34	0.53
Truck	A2	B2	C2	D2	E2
Effect	0.14	−0.12	−0.28	−0.28	−0.02

Table 6. Estimates of the variance components of the factors in the mixed effects model and in the random effects model (from Equations (4) and (5)). They were obtained through the (restricted) maximum likelihood procedure. We also included the estimates of the standard deviations of the factors by taking the square roots of the estimated variances.

Factor	Variance component	Standard deviation	Percentage of total variance
Route	0.1147	0.3386	9.7%
Day	0.6511	0.8069	55.0%
Truck	0.0156	0.1251	1.3%
Error	0.4018	0.6338	34.0%

The estimated fixed effects for days are the same as the least-square estimates given in Table 5. Table 6 also contains the estimated variance parameter, σ_D^2, under the random effects model to be discussed in Section 4.3. This should be ignored when thinking of the mixed model parameters.

4.3. Average Load and Truck Capacities Under the Random Effects Model

The random effects model differs from the fixed effects model in assigning a distribution to the day effects. While days are not actually sampled from a distribution, it is nonetheless useful to compare how the estimated variance component for days differs from those obtained for truck and route.

Formally, the mixed effects model can be converted into a random effects model by assigning a distribution to the day effects and augmenting Equation (4) with an additional component given by

$$\tau_d \sim N\left(0, \sigma_D^2\right), \quad d = 1, 2, \ldots, 5. \tag{5}$$

The model defined by Equations (1), (3), (4), and (5) is the random effects model. We have assumed independence in our assigned distributions for each of the factors and the variance of a load, y_{rdt}, is given by the sum of the variances of the constituent random variables in (1):

$$\mathrm{var}(y_{rdt}) = \sigma_R^2 + \sigma_D^2 + \sigma_T^2 + \sigma^2. \tag{6}$$

Equation (6) provides a decomposition of the total variation into four distinct components and allows one to ask for the relative importance of each of the error terms.

Table 6 displays the estimated variance components of the three factors. These estimates for the factors of route and truck are the same in the random effects model as in the mixed effects model. When there are no interactions, the two models will have the same estimates for the variances of the random components. This is because the decomposition of the total sum of squares into these component sum of squares values is the same in a mixed effects model as it is in a random effects model (for any given dataset). However, this will not be the case if there are interactions in the model, because they will affect the decomposition of the sum of squares.

Note that the truck factor has a variance of 0.0156, which is the smallest contribution to differences in loads. Also shown are the square roots of the variance, the estimated standard deviations of the effects, which are more easily interpretable. From Table 6, the estimated truck effects standard deviation, 0.1251 tons, is the smallest of the three. The route effects standard deviation, 0.3386, is almost three times that of truck, while the day standard deviation, 0.8069, is over six times that of truck.

To test for the presence of the truck effect, we perform a likelihood ratio test. In this test, we compare the “reduced model” (without the truck effect) to the “current model” (with the truck effect). Small p-values (below the standard threshold of 0.05 or 0.01) provide evidence against the reduced model (without the truck effect.) We obtain a p-value of 0.11, which is larger than any reasonable threshold, so there is not enough evidence against the reduced model. In other words, the likelihood ratio test gives no evidence that there is a truck effect.

The square root of the residual variance, at 0.6338 tons, is sizable. The measurement error in weighing the loads is believed to be a very small fraction of this amount. This term represents the collective effects not explained by the model. Although the size of the unexplained component is significant, the purpose of the experiment was to study the performance of the trucks under typical variations to which they are exposed and allow us to study the grand mean and the true truck means.

In reporting the random effects model, having variance components for all factors—including day—we take the point of view proposed by Gelman (2005, p. 2), in which variance components in a random effects model are viewed in a descriptive or exploratory way. He writes: “Our approach is to use variance components modeling for all rows of the table [as here in Table 6], even for those sources of variation that have commonly been regarded as fixed effects;” and Gelman (2005, p. 33, caption to Fig. 5): “Compared to the classical ANOVA …this display makes apparent the magnitudes and uncertainties of the different components of variation.” This approach exemplifies the

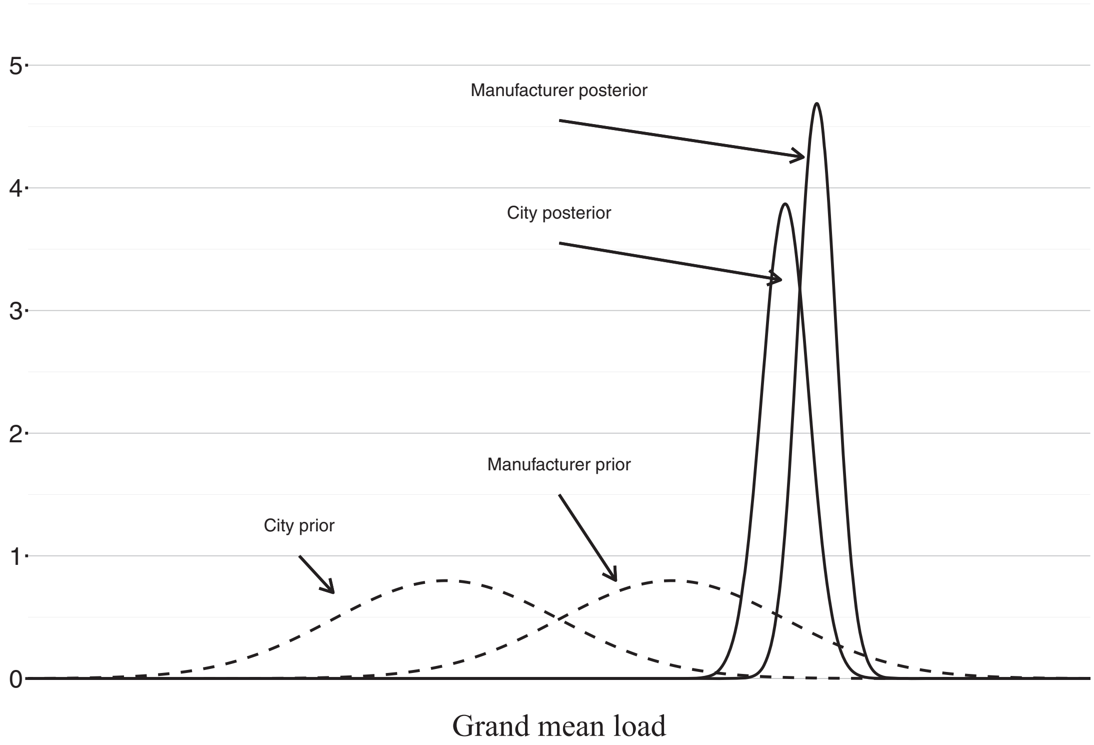

Figure 5. Plotted are the prior densities and the posterior densities of the grand mean of loads of the city and of the manufacturer. Data from the test bring the means of the posterior densities (solid lines) for the grand mean load, μ, for the city and the manufacturer closer together than the means of their prior densities for μ (dashed lines). Both posterior means are well above the specification of 12.5 tons.

view expressed by John Tukey " …that what is a fixed or a random effect is best not answered as if these were intrinsic properties of the phenomena under study, but rather that the questions to be answered should guide the choice" (see Robinson 1991, p. 29, quoting Tukey).

We now turn away from the frequentist perspective and consider fitting a model from a Bayesian viewpoint, which is the topic of Section 5.

5. A Bayesian Perspective

The city was initially skeptical that the trucks would meet the specification, while the manufacturer was confident that they would. Their differing views can be described with different prior distributions on the expected number of tons a truck will load. In this section, we develop this Bayesian perspective on the dispute. In reality, the parties did not proceed armed with priors nor were they shown their posteriors for the grand mean. However, we think it is interesting to see how such a perspective might work in a dispute, using the dispute recounted in this article as a case study.

5.1. *The City's Prior and the Manufacturer's Prior on the Grand Mean*

Assume that the mean of the city's prior on the grand mean load, μ, *does not* satisfy the contract specification of 12.5 tons (for illustration, assume the city's prior mean is 12 tons) while the mean of the manufacturer's prior *does* (for illustration, the manufacturer has a prior mean of 13). If the parties are conducting the test in good faith, arguably they will want to choose variances for their prior distributions for μ that assign some nonnegligible probability to outcomes that are different from their own initial expectations and that include the expectations of the other party. A standard deviation of 0.5—given in Section 5.2—for both the city's and the manufacturer's priors on the grand mean load, μ, in the model (1) puts the parties' priors 2 standard deviations apart (13-12 = 1 = 2*0.5), which satisfies the good faith criterion. Section 5.2 gives the details of the calculations of the posteriors. Section 5.3 gives the results.

5.2. *A Bayesian Fixed Effects Model of Loads*

We implement the fixed effects model from a Bayesian perspective in this section. We focus on fixed effects because this leads to closed-form results. It would be interesting to develop a mixed and random effects model from a Bayesian perspective as well. However, since these would not lead to analytical results, we would need an application of a Markov chain Monte Carlo sampler, which goes beyond the scope of this article.

Letting $\boldsymbol{\beta}$ denote a vector of 18 independent parameters (1 for the grand mean, 4 for route effects, 4 for day effects, 9 for truck effects), the specification for the prior for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{\mu_0}, \sigma^2\boldsymbol{\Lambda}_\mathbf{0}^{-\mathbf{1}}\right), \qquad (7)$$

where $\boldsymbol{\mu_0}$ is an 18-element prior mean, and $\sigma^2\boldsymbol{\Lambda}_\mathbf{0}^{-\mathbf{1}}$ is an 18x18 prior covariance matrix of the parameters $\boldsymbol{\beta}$.

Table 7. Prior mean and standard deviation of the parameters of the mean model at Equation (1).

Factor	Number of parameters	City prior mean (tons)	Manufacturer prior mean (tons)	Standard deviation (tons)
Grand mean	1	12	13	0.50
Route	4	0	0	1.00
Day	4	0	0	1.00
Truck	9	0	0	1.00
Total	18			

The diagonal elements of $\sigma^2\boldsymbol{\Lambda}_\mathbf{0}^{-\mathbf{1}}$ represent how confident the practitioner is in their prior information on the elements of the $\boldsymbol{\beta}$ vector, that is, the grand mean, route, day, and truck. Smaller variances correspond to higher confidence.

In the implementation, we take the city's prior for the mean parameter vector $\boldsymbol{\beta}$ to be a vector whose first component is 12 tons, and whose remaining 17 components are 0; and we take the manufacturer's prior to be the same vector, except that the first component is 13 tons. Table 7 summarizes the choices of prior means and prior standard deviations for the mean model for tons loaded defined at Equation (1).

The rationale for the prior standard deviations on the mean parameters are as follows. The standard deviation for the grand mean is 0.5 tons. With this choice for the prior, the city and the manufacturer each assign nonnegligible probability to outcomes that are different from their own initial expectations, that is, their prior means of 12 and 13 tons, respectively, and that include the expectations of the other party. The standard deviation for the route effects is 1 ton. This choice reflects a belief that the City Sanitation Department would know if much more variation than that was plausible. The standard deviation for the day effects is 1 ton. This choice similarly reflects a belief that the City Sanitation Department would know if much more variation than that was plausible. The standard deviation for the truck effects is 1 ton. The better the manufacturing tolerances for truck hydraulics, the smaller this value could be.

If we choose a distribution for $\boldsymbol{\beta}$ and σ^2 that is conjugate to the distribution of y_{rdt}, given $\boldsymbol{\beta}$ and σ^2, then the posterior distribution of $\boldsymbol{\beta}$ and σ^2 given y_{rdt} will be of the same family as the prior. The following prior for σ^2 makes $\boldsymbol{\beta}$ and σ^2 conjugate in this way:

$$\sigma^2 \sim \text{Inv-Gamma}\,(a_0, b_0)\,. \qquad (8)$$

To make it uninformative, we take the prior parameters for the inverse gamma distribution at (8) to be $a_0 = b_0 = 1$ for both the city and the manufacturer. See Gelman et al. (2014, pp. 42–43). Then the posterior distribution for the unknown parameters is

$$\begin{aligned}
[\boldsymbol{\beta}|\boldsymbol{Y},\sigma^2] &\sim N((\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{\Lambda_0})^{-1}(\boldsymbol{X}^T\boldsymbol{Y}+\boldsymbol{\Lambda_0}\boldsymbol{\mu_0}),\\
&\qquad (\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{\Lambda_0})^{-1}\sigma^2) \text{ and}\\
[\sigma^2|\boldsymbol{Y}] &\sim \text{Inv-Gamma}\left(a_0 + n/2,\, b_0\right.\\
&\qquad \left. + \frac{1}{2}\left(\boldsymbol{Y}^T\boldsymbol{Y} + \boldsymbol{\mu_0}^T\boldsymbol{\Lambda_0}\boldsymbol{\mu_0} - \boldsymbol{\mu_n}^T\boldsymbol{\Lambda_n}\boldsymbol{\mu_n}\right)\right), \qquad (9)
\end{aligned}$$

where $\boldsymbol{\mu_n} = \boldsymbol{\Lambda_n}^{-1}(\boldsymbol{X}^T\boldsymbol{Y}+\boldsymbol{\Lambda_0}\boldsymbol{\mu_0})$ and $\boldsymbol{\Lambda_n} = \boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{\Lambda_0}$. We implement the model by taking 1000 draws of σ^2, and using those values to then draw 1000 values of $\boldsymbol{\beta}$. This will give us draws from the joint posterior distribution of $\boldsymbol{\beta}$ and σ^2.

5.3. Comparison of the Two Posteriors

The data from the test, whose mean is 13.74 tons, pull both the city's and the manufacturer's posterior densities toward it, and pull the two posteriors closer together than the priors. Figure 5 shows the city's and the manufacturer's prior densities for the mean load, μ (dotted lines), and their respective posterior densities (solid lines).

The city's posterior density for μ has a mean of 13.61 tons, while the manufacturer's posterior density for μ has a mean of 13.69 tons. The posterior densities show how the data have brought the initial beliefs of the parties much closer together—from a difference between the prior means of 1 ton ($1 = 13 - 12$) to a difference between the posterior means of 0.08 tons ($0.08 = 13.69 - 13.61$); both posterior mean loads are substantially greater than the specification of 12.5 tons.

The data information is 50 (the number of observations), and the prior information is 4 (the inverse of the prior variance of 0.25 on the grand mean). Therefore, the data weight is 50/54 $= 0.925926$ and the prior weight is $4/54 = 0.074074$. Applying these weights to the data mean and prior means, respectively, gives the posterior means of 13.61 and 13.69 for city and manufacturer, respectively, as noted in the previous paragraph. Thus, the posterior means are much closer to the data mean of 13.74 tons than to their prior values.

Under the Bayesian model with prior parameters given in Table 7, the means of the distributions of the effects of the three factors are all shrunk toward 0 in comparison to their values in the data, that is, given by the deviations of the factor level load averages from the grand load average. The shrinkage reflects their prior means of 0 and the informative prior standard deviations. For example, and of most interest to the parties, the 10 truck effects in the data are, from Table 5, in order of size: −0.40, −0.28, −0.28, −0.16, −0.12, −0.02, 0.14, 0.26, 0.34, 0.53—having a range of 0.93 tons. The means of the posteriors of effects are: −0.38, −0.27, −0.26, −0.15, −0.12, 0.00, 0.14, 0.24, 0.32, 0.50—having a range of 0.88 tons, and corresponding to a shrinkage of 5%.

The shrinkage is sensitive to the sizes of the standard deviations assumed for the prior on the 18 model parameters. For example, if, in place of the values given in Table 7, all the standard deviations were doubled, creating a less informative prior, the range of the posterior means of the truck effects (using the city's prior), at 0.89 tons, or a 4% decrease from 0.93, is a decreased shrinkage. As the standard deviations are increased to infinity, the shrinkage disappears. On the other hand, if all the standard deviations were halved, creating a more informative prior, the range, at 0.82, or a 12% decrease from 0.93, is an increased shrinkage.

Practically speaking, for the issue at hand, shrinkage means that differences between trucks are smaller than the data alone would indicate, that is, the trucks are more reliable. Closely similar results hold using the manufacturer's prior. See Box and Tiao (1968) for another example of shrinkage of means in an analysis of variance model.

We can also look at the posterior probability that loads will meet the specification on each route and on each day—for both the city's and manufacturer's prior. The city has an interest in seeing that, given the test results, there is a reasonable probability that the trucks will perform satisfactorily regardless of the environment, that is, regardless of route and day.

Table 8. Posterior probabilities for the first set of five trucks that loads will be 12.5 or more tons under the city's prior of 12 tons for the grand mean. Results for the second set of five trucks, and using the manufacturer's prior are quite similar. Probabilities are computed from the posterior predictive distribution, Equation (10).

	Day				
Route	M	T	W	TH	F
1	0.837	0.993	0.790	0.835	0.637
2	0.997	1.00	0.924	0.941	0.756
3	0.971	0.998	0.942	0.677	0.412
4	0.943	0.969	0.857	0.415	0.641
5	0.953	0.987	0.976	0.890	0.761

To see how trucks will perform in the future, we must draw from the posterior distribution for a previously unobserved truck (or the same truck taking a new run), $\tilde{y}_{rdt}$, that follows the same process as the trucks we have observed in our dataset. This is called the posterior predictive distribution.

$$[\tilde{y}_{rdt}|\boldsymbol{Y}] \sim t\left(2a_0 + n; \boldsymbol{x}_{rdt}^T\boldsymbol{\mu}_n; (2b_0 + \boldsymbol{Y}^T\boldsymbol{Y} + \boldsymbol{\mu}_0^T\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 \right. \\ \left. - \boldsymbol{\mu}_n^T\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n)\left(1 + \boldsymbol{x}_{rdt}^T\boldsymbol{\Lambda}_n^{-1}\boldsymbol{x}_{rdt}\right)/(2a+n)\right), \quad (10)$$

where $2a_0 + n$ are the degrees of freedom, $\boldsymbol{x}_{rdt}^T\boldsymbol{\mu}_n$ is the center, and $(2b_0 + \boldsymbol{Y}^T\boldsymbol{Y} + \boldsymbol{\mu}_0^T\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^T\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n)(1 + \boldsymbol{x}_{rdt}^T\boldsymbol{\Lambda}_n^{-1}\boldsymbol{x}_{rdt})/(2a+n)$ is the scale. $\boldsymbol{x}_{rdt}$ is the coefficient vector of $\boldsymbol{\beta}$ that identifies the rth route, dth day, and tth truck. See Gelman (2005) for derivation.

Table 8 gives these posterior probabilities. Under either the city's or the manufacturer's prior for the grand mean, the large majority of route and day combinations have probabilities of reaching or exceeding the specification that are above 0.8, though there are two that are below 0.5.

6. Discussion

In the end, the city conceded that the test had demonstrated the fix to the hydraulics offered by the manufacturer had worked. The city accepted delivery of the order and the manufacturer was paid. The Latin Square design was instrumental in understanding the capacities of the trucks under different operating conditions. It allowed one to disentangle the effects of the environment in which the truck would operate from the performance of the trucks themselves.

The design of the test and the statistical treatment that followed facilitated a resolution of the dispute between the city and the manufacturer. The test showed that the broad distribution of measured loads, including six that fell below the contract specification of 12.5 tons, was not due to differences between trucks, but rather it was due to the different environments in which the trucks were operated—namely, day of the week and the type of route on which they were run. The mean load of the truck runs, 13.74 tons, was well in excess of the 12.5 ton requirement, and the difference between the estimated mean and the contract specification was statistically significant.

A Bayesian analysis that uses a prior distribution centered at 12 tons for the city, and a prior centered at 13 tons for the manufacturer, is presented as a way in which to understand how two sides with differing initial opinions might revise these views in light of the data obtained. The Bayesian approach might be developed into a useful tool for negotiations. It embraces the fact that the parties may have divergent views, and it provides a method for either or both parties to modify these views given new information. In commercial disputes, parties are accustomed to thinking in terms of probabilities. The law itself frames questions in terms of chance, for example, the standard of proof in civil cases is whether the facts presented make the question under examination "more likely than not." (see Kaye 1982, p. 515). This approach is similar in spirt to the Bayesian perspective offered here, which may prove to be a helpful guide in resolving disputes of this nature.

Several aspects of this case study have been used as examples in at least four courses in two universities. Students have generally been receptive, often commenting that the real life example provides them motivation to more fully understand a number of the concepts involved and to delve more deeply into assumptions and modeling choices that are made when performing a real analysis.

This case study is suitable for use in a second course on statistics, in which students have already obtained some understanding of the methods and approaches one might apply, a course on regression analysis or linear models, or a course on Bayesian statistics. For a second course in applied statistics, it provides an example of a realistic experiment, of the use of a Latin Square to obtain a balanced design—allowing the use of far fewer observations than would be required in a full factorial experiment, of sources of variability, of fixed and random effects, of linear statistical models, of three-way ANOVA, and of Bayesian inference. This material can also be used in discussions of the role of a consulting statistician, of examples of the use of statistics in negotiations and the law, and of practical concerns with which one is faced when performing a real experiment.

Supplementary Material

The data, "truck_data2.csv," and "R Code for Sections 1-4," as well as "R Code for Section 5 - A Bayesian Perspective" for this article are available online.

Acknowledgments

The authors thank Anteneh Tesfaye and Barbara Donnini for assistance with computing; Edward Cheng, Philip Everson, Marian Farah, and Judith Tanur for comments; thanks to anonymous referees for their highly useful comments; and City officials for their enlightened use of statistical methods in the actual case.

References

Box, G. E. P., and Tiao, G. C. (1968), "Bayesian Estimation of Means for the Random Effects Model," *Journal of the American Statistical Association*, 63, 174–181. [257]

Casella, G. (2008), *Statistical Design*, New York: Springer. [250]

Gelman, A. (2005), "Analysis of Variance—Why it is More Important than Ever," *Annals of Statistics*, 33, 1–53. [253,255,257]

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014), *Bayesian Data Analysis*, Boca Raton, FL: CRC Press. [256]

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., eds. (1991), *Fundamentals of Exploratory Analysis of Variance*, New York: Wiley. [252]

Juran, J. M., and Godfrey, A. B. (1988), *Juran's Quality Handbook* (5th ed.), New York: McGraw–Hill. [249]

Kaye, D. (1982), "The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation," *American Bar Foundation Research Journal*, 7, 487–516. [258]

Robinson, G. K. (1991), "That BLUP is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6, 15–32. [256]