Focus Article

Applications of subset selection procedures and Bayesian ranking methods in analysis of traffic fatality data

Gary C. McDonald*

Nonparametric and parametric subset selection procedures are used in the analysis of state motor vehicle traffic fatality rates (MVTFRs), for the years 1994 through 2012, to identify subsets of states that contain the 'best' (lowest MVTFR) and 'worst' (highest MVTFR) states with a prescribed probability. A new Bayesian model is developed and applied to the traffic fatality data and the results contrasted to those obtained with the subset selection procedures. All analyses are applied within the context of a two-way block design. © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

How to cite this article:
WIREs Comput Stat 2016, 8:222–237. doi: 10.1002/wics.1385

Keywords: Fatality Analysis Reporting System, probability of a correct selection, Bayesian inference, WinBugs, additive model, Tukey one-degree-of-freedom test for additivity

INTRODUCTION

This article focuses on the application of nonparametric (or distribution-free) and parametric subset selection procedures to analyze motor vehicle traffic fatality rate (MVTFR) data for the years 1994 through 2012. As such, much of the data analysis done in this study builds upon the work of Green and McDonald[1] analyzing MVTFR for the years 1982 through 2002 as well as several earlier studies. In that study, the states selected to contain the worst state consist primarily of the Southeastern states and several states in the Northwest and Southwest. The states selected to contain the best state mostly include states along the East Coast, selected North Central States, and the state of Washington.

In addition to applying the selection procedures to an updated dataset, this article also presents and applies a new Bayesian approach to the ranking of states. With this approach, a probability distribution is derived over all possible permutations of the population means. Thus, the probability that any particular state is characterized by the largest (or smallest) mean can be easily obtained by appropriate summing of the permutation probabilities.

FORMULATION OF NONPARAMETRIC SUBSET SELECTION RULES

The description of this selection rule will follow that given by Green and McDonald.[1] Let $\Pi_1,\Pi_2,\ldots,\Pi_k$ be k ($\geq$2) independent populations. The associated random variables, X_{ij}, $j = 1,\ldots,n$; $i = 1,\ldots,k$, are assumed independent and to have a continuous distribution $F_j(x;\theta_i)$ where θ_i belong to some interval Θ on the real line. The basic model assumption is that $F_j(x;\theta)$ is a stochastically increasing family of distributions for each j. The additive model of the following form is used:

*Correspondence to: mcdonald@oakland.edu

Department of Mathematics and Statistics, Oakland University, Rochester, MI, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

© 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

$$X_{ij} = \mu + \theta_i + \beta_j + \varepsilon_{ij} \tag{1}$$

where β_j indicates the particular block effect, θ_i indicates the population effect, and ε_{ij} is the random error. The distribution of ε_{ij} is any continuous distribution function $F_j(x)$ with mean 0. The distribution of X_{ij} will be stochastically ordered in θ as it is a location parameter in Eq. (1). So, for example, $F_j(x)$ could be a normal distribution with mean 0 and standard deviation σ_j. The assumption of negligible interaction between population and block must be satisfied. Let $\theta_{[i]}$ denote the ith smallest unknown parameter, then for all x

$$F_j(x;\theta_{[1]}) \geq F_j(x;\theta_{[2]}) \geq \ldots \geq F_j(x;\theta_{[k]}) \tag{2}$$

where $\theta_{[1]}$ ($\theta_{[k]}$) characterizes the best (worst) population.

Let R_{ij} denote the rank of the observation X_{ij} among $X_{1j},X_{2j},\ldots,X_{kj}$. The variables R_{ij} take values from 1 to k. The selection procedures considered here are based on the rank sums, $T_i = \sum_j R_{ij}$, associated with Π_i, $i = 1,\ldots,k$. The structure for this process is outlined in Table 1.

Any subset selection procedure based on the rank sums should have the property that the probability that a correct selection (CS) occurs, i.e., the worst population (or best population) is included in the selected subset, is bounded below by P^* ($k^{-1} < P^* < 1$). That is, for a given selection rule R, the probability of a CS should satisfy the inequality,

$$\inf_{\Omega} P(\mathrm{CS}|\mathrm{R}) \geq P^*, \tag{3}$$

where $\Omega = \{\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k): \theta_i \in \Theta, i = 1,\ldots,k\}$. In some cases, as noted later, this inequality may only hold on a subspace Ω' of Ω.

The two selection rules for choosing a subset containing the worst population, as described in McDonald,[2] are given by:

R_1: Select Π_i iff $T_i \geq \max(T_j) - b_1$
R_2: Select Π_i iff $T_i > b_2$.

Similarly, the two selection rules for choosing a subset containing the best population are given by:

R_3: Select Π_i iff $T_i \leq \min(T_j) + b_3$
R_4: Select Π_i iff $T_i < b_4$.

Note that the rules R_1 and R_2 could be written in the form that select Π_i iff $T_i \geq b$, where b is a stochastic quantity for R_1 and a deterministic quantity for R_2. A similar statement can be made for the rules R_3 and R_4.

As developed by McDonald,[3–5] R_1 and R_3 are justified over a slippage space, Ω', where all parameters θ_i are equal with the possible exception of $\theta_{[k]}$ in case of rule R_1 or $\theta_{[1]}$ in case of rule R_3; and R_2 and R_4 are applicable over the entire parameter space. The constants b_1, b_3, and b_4 are chosen as small as possible and b_2 is chosen as large as possible preserving the probability goal. For large values of n, the selection rules are determined by the asymptotic formulae as described in McDonald[4] and are computed as:

$$b_1 = b_3 = h[nk(k+1)/6]^{1/2}, \tag{4}$$

$$b_2 = \left[n\left(k^2-1\right)/12\right]^{1/2}\Phi^{-1}(1-P^*) + n(k+1)/2, \tag{5}$$

$$b_4 = n(k+1) - b_2, \tag{6}$$

where the h-solution to be used in Eq. (4) is given by:

$$\int_{-\infty}^{\infty} \Phi^{k-1}\left(x + h2^{1/2}\right)\phi(x)\mathrm{d}x = P^*. \tag{7}$$

Here, Φ and ϕ represent the standard normal cumulative distribution function (CDF) and probability density function (PDF), respectively.

Taking P^* to be a particular confidence level, the h-solution is given in Table 1 of Gupta et al.[6] and can be used to determine the constants b_1 and b_3. The above integral can also be calculated to

TABLE 1 | Structure for Determining Ranks and Rank Sums

Block/Π	Π_1	Π_2	···	Π_k	SUM
Block 1	$X_{11} \approx R_{11}$			$X_{k1} \approx R_{k1}$	$k(k+1)/2$
Block 2	$X_{12} \approx R_{12}$			$X_{k2} \approx R_{k2}$	$k(k+1)/2$
.	.			.	.
.	.			.	.
.	.			.	.
Block n	$X_{1n} \approx R_{1n}$			$X_{kn} \approx R_{kn}$	$k(k+1)/2$
RANK SUMS (T_i)	T_1			T_k	$nk(k+1)/2$

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

determine P^* for a given value of h, using a TI-83+ (or similar) calculator with numerical integration capability as shown in Green and McDonald.[1] The integral can be shown to be the probability that the maximum of U_i, $i = 1,\ldots,k$, is less than h where the U_i are normally distributed random variables with zero means, unit variances, and covariance of 0.5. With confidence level P^*, it can be asserted (using these selection rules) that the chosen subset of the populations contains the one characterized by $\theta_{[k]}$ ($\theta_{[1]}$).

AN APPLICATION TO STATE MOTOR VEHICLE TRAFFIC FATALITY RATES

The analysis of MVTFRs described in this section follows the approach taken by Green and McDonald.[1] The database herein used has been updated by 8 years from that used in 2009. This analysis is included in this article to illustrate the use of the nonparametric selection procedures, to update a statistical analysis, and to set the framework for the application of parametric rules and a new Bayesian approach to this class of problems. The methods herein applied to MVTFRs provide statistical analyses of the data yielding probabilistic guarantees of inference for a specific class of ranking questions complementing other descriptive techniques, e.g., see Sivak.[7]

The scope of this study includes highway MVTFR data for the 51 U.S. states (treating the District of Columbia as a state) from year 1994 through 2012. The National Highway Traffic Safety Administration (NHTSA) publishes the MVTFRs for all U.S. states each year in the Fatality Analysis Reporting System (FARS). The data can be accessed through the government website: www-fars.nhtsa.dot.gov. The data are included here as Appendix A. The FARS Encyclopedia provides extensive, detailed statistics on injuries and deaths from vehicle accidents that occurred within the 50 states and the District of Columbia. The fatality rate per year for each state is expressed as the number of fatalities per 100 million vehicle miles of travel (VMT). It should be noted that when new exposure data are released, the previous years' exposure data is updated. Appendix A lists data posted in March, 2015. Current listing (March, 2016) includes rates for 2013 and, hence, some changes in the 2012 rates shown in Appendix A.

The model used is that of the additive form (1), where θ_i is the state effect, β_j is the year effect, and ε_{ij} is the random error term. This two-way model is used in earlier studies of MVTFRs by McDonald,[2] by Lorenzen and McDonald,[8] by Green et al.,[9] and by Green and McDonald.[1] It is discussed extensively by Neter et al.,[10] Kuehl,[11] Winer,[12] Christensen,[13] and many others. The ε_{ij} may have any (not necessarily normal) continuous distribution. The observations are taken in n blocks (time periods). The subscript j indicates the particular block (year) to which the observation X_{ij} corresponds and i indicates the population (state).

Figure 1 provides boxplots for the MVTFR data indicating visually a substantial effect of the variables year and state. The boxplots on the left side of Figure 1 shows the distributions of the MVTFRs of all states decreasing for the years 1994, 2003, and 2012. The median rates for 1994 and 2012 are

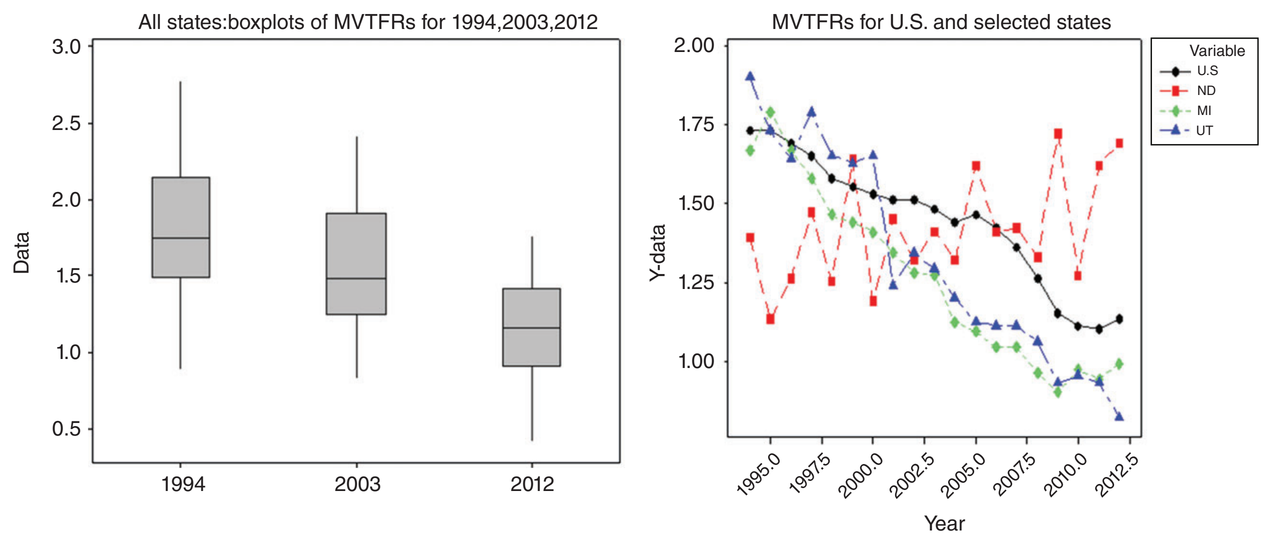


FIGURE 1 | Boxplots of MVTFRs. MVTFRs, motor vehicle traffic fatality rates.

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

approximately 1.7 and 1.2, respectively, an almost 30% decrease. The scatter plot on the right side of Figure 1 shows the MVTFRs of the states North Dakota (ND), Michigan (MI), and Utah (UT) along with the overall U.S. rates for the years 1994–2012. While the overall trend is strongly decreasing, not all states depict the same behavior as demonstrated by that of ND.

Since there is only one observation for each state for each year, there is no general test for additivity, i.e., lack of interaction between states and years. Tukey[14] developed a one degree-of-freedom test for nonadditivity when there is a single observation per cell, as given here. Green et al.[9] and Green and McDonald[1] use this test to establish the plausibility of model (1) for a power transformation of the MVTFRs. Table 2 shows the Tukey one degree-of-freedom test for nonadditivity for the MVTFRs and for these rates raised to the 0.3 power. The test indicates significant evidence of interaction with the untransformed rates, and no significant evidence of interaction with the power transformation of the rates. For the purpose of the nonparametric analyses to follow, the original MVTFR data will be used because ranks are invariant to monotone increasing transformations.

NONPARAMETRIC SUBSET SELECTION OF STATES

The goal is now to choose a subset of the 51 states that can be asserted, with a specified confidence, to contain the state with the highest MVTFR (worst population), and similarly a state with the lowest MVTFR (best population) using the nonparametric ranking and selection procedures. Ranks k = 1, … , 51 are assigned to states for each of n = 19 years, with a rank of '1' being the state with the lowest MVTFR. Based on these ranks, the selection procedure for choosing a subset of the 51 states asserts that the best state (or worst state) is contained with a specified confidence level P^*.

TABLE 2 | Tukey's One Degree-Of-Freedom Test

• Tukey's 1 DF Test of Nonadditivity—MVTFR	• Tukey's 1 DF Test of Nonadditivity—$MVTFR^{0.3}$
• SS(Nonadditivity): 1.661	• SS(Nonadditivity): 0.004
• SS(Error): 17.482	• SS(Error): 0.931
• MS(Error): 0.019	• MS(Error): 0.001
• Significance level: 0.050	• Significance level: 0.050
• Test statistic: 85.419	• Test statistic: 3.384
• Critical value: 3.852	• Critical value: 3.852
• The test statistic is greater than the critical value, so there is significant evidence of interaction.	• The test statistic is not greater than the critical value, so there is no significant evidence of interaction.

MVTFR, motor vehicle traffic fatality rates; SS, sum of squares; MS, mean square.

Similar to the structure as outlined in the second section, let R_{ij} denote the rank of the observation X_{ij} within the jth block. The variables R_{ij} take values from 1 to k and the selection procedure is based on the rank sums, $T_i = \Sigma_j R_{ij}$, associated with Π_i, $i = 1, \ldots, k$. For the first year, 1994, Rhode Island has a rank of '1' and the state of Mississippi has a rank of '51.' In the case of ties, each tied state receives an average of their rank for that year. This is done for all 19 years. Ranks are then summed for each state and the rank sums, T_i's, are ordered. Since the values of k and n are large for our application (k = 51, n = 19), the selection rule constants are determined by the asymptotic formulae as described in the second section.

Taking $P^* = 0.90$, the h-solution as given in Table 1 of Gupta et al.[6] is h = 2.5920. This can be used to determine the constants b_1 and b_3. Using n = 19, k = 51, and h = 2.5920, we obtain $b_1 = b_3 = 237.532$. The other two constants are calculated to be $b_2 = 411.774$ and $b_4 = 576.226$. The data yields $\max(T_j) = 930.5$ and $\min(T_j) = 23$.

With confidence level $P^* = 0.90$, it can be asserted that the following subsets of states contain that one characterized by $\theta_{[k]}$:

Rule R_1: Select the ith state iff $T_i \geq \max(T_j) - 237.532 = 692.968$. Sixteen are chosen for 'worst.'

Rule R_2: Select the ith state iff $T_i > 411.774$. Thirty are chosen for 'worst.'

With the same 0.90 confidence level, it can be asserted that the following subset of states contain that one characterized by $\theta_{[1]}$:

Rule R_3: Select the ith state iff $T_i \leq \min(T_j) + 237.532 = 260.532$. Twelve are chosen for 'best.'

Rule R_4: Select the ith state iff $T_i < 576.226$. Twenty nine are chosen for 'best.'

The identification of the specific states chosen with these four selection rules is given in Appendix B. The states chosen by rules R_1 and R_3 are graphically displayed in Figure 2. The 16 states chosen by R_1 to contain the 'worst' state consist primarily of the Southeastern states and some states in the Northwest and Southwest. The 12 states chosen by R_3 to contain the 'best' state mostly include states from the Northeast along with Minnesota (MN), Washington (WA), Virginia (VA), and California (CA). The inference conditions have been discussed earlier.

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

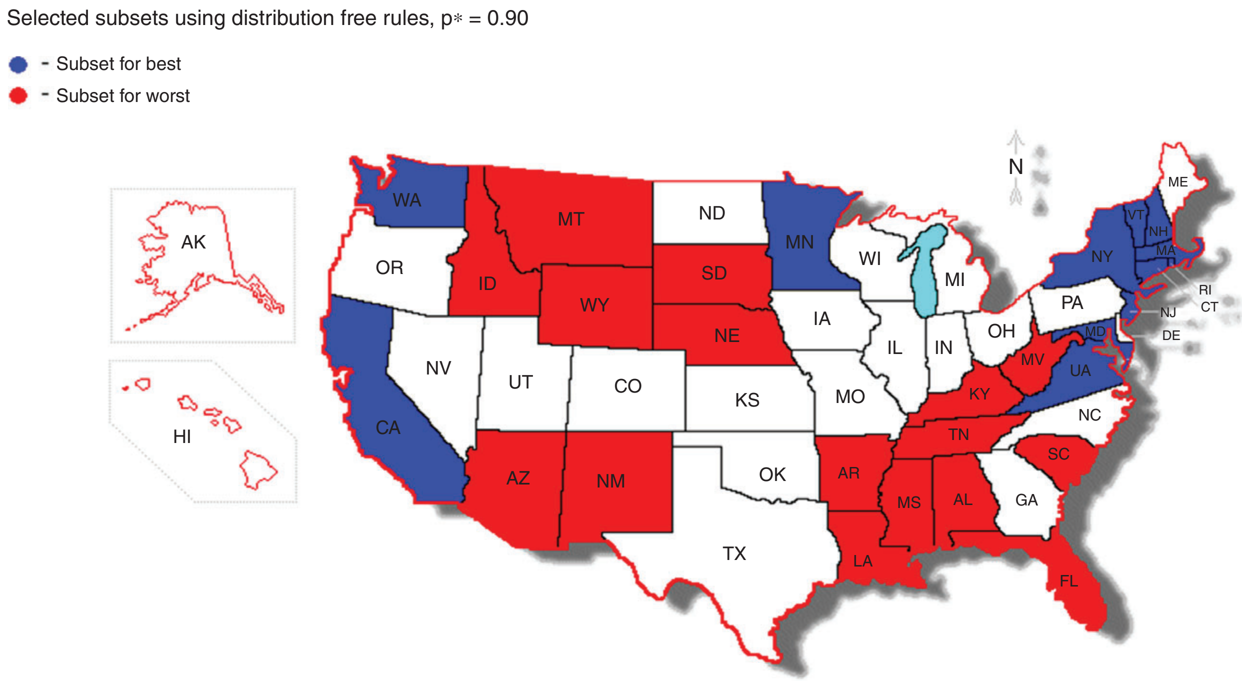


FIGURE 2 | States selected using rules R_1 and R_3.

These results suggest that the particular separation of states is attributable to the differences among states in the urban, rural, and interstate mileage mix. This hypothesis is examined for earlier traffic fatality data by Lorenzen and McDonald.[8] The authors apply these nonparametric selection rules to state traffic fatality rates adjusted for the state urban/rural mileages. Two adjustment methods commonly used on census data and on mortality data, the direct method and the indirect method, are discussed in detail in this reference. With the direct method, the actual rural and urban rates of each state are weighted by the same urban/rural mix, typically the national average mix.

The indirect method is slightly more complex. With it the national fatality rate is multiplied by the ratio of the state's actual fatality rate divided by the state's expected fatality rate calculated with the state's actual urban/rural mix and some specified (usually national) urban and rural fatality rates. Detailed references on the direct and the indirect method are Woolsey,[15] Duffy and Carroll,[16] and Bishop et al.[17] While the two adjustments, in practice, may give similar results, each method has specific properties. The Lorenzen and McDonald[8] article discusses extensively the data sources available for adjusting rates for these two methods, applies the nonparametric selection rules to the adjusted rates, and summarizes extensive Monte Carlo simulations quantifying the inference properties for the nonparametric statistical procedures. Other plausible causes for the observed separation of states are also noted.

The choice of a fatality rate index can be impactful. While not investigated in this article, a somewhat related article by Gibbons and McDonald[18] examined the sensitivity of an air pollution and health study to the choice of a mortality index. In that study, the authors examined the sensitivity of conclusions based on regression models to the mortality index incorporated as the dependent variable. Four indices were examined, including direct and indirect adjusted rates.

The validity of the nonparametric analysis herein used depends on the legitimacy of the additive model (1). An argument, based on the Tukey test for nonadditivity, has been given to justify the plausibility of it. However, other forms of interaction could be present in such a manner that time impact differs among states in a nonadditive fashion. To partially address this issue, an additional analysis was conducted on two time segments of the data. The first segment included years from 1995 to 2003, and the second segment from 2004 to 2012. The year 1994 was not included here so as to have equal sample sizes ($n = 9$) for the two periods. Selection rules R_1 and R_3 were applied to these two datasets with $P^* = 0.90$ and the selection rule constants b_1 and b_3 computed from Eq. (4). The Pearson correlation coefficient for the state rank sums for these two time

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

periods is 0.922 indicating a rather close linear relationship.

Applying R_1 to the two datasets, 16 states are included in the 'worst' subset for each of the two time periods. These 16 states include 15 of the 16 states, noted in Appendix B, chosen when R_1 was applied to the complete dataset. Applying R_3 to the two datasets, 17 states are included in the 'best' subset for each of the two time periods. These 17 states include the 12 states, noted in Appendix B, chosen when R_3 was applied to the complete dataset. Notably, the analysis of the first time period fatality rates placed ND in the 'best' subset. However, the analysis of the second time period fatality rates placed ND in the 'worst' subset. In the analysis of the combined data, ND was not chosen by either rule R_1 or R_3. A comprehensive detailed analysis of the ND crash data is published by the North Dakota Department of Transportation.[19]

As observed in the application of selection rules R_1 and R_2, the rule R_2 always selects more populations than R_1. This appears natural as the inference properties for R_2 are more expansive than those established for R_1. However, Green and McDonald[1] provide a counterexample showing that R_1 can result in a larger subset than R_2. To further address this property, note that R_2 *always* selects at least as many populations as R_1 if $b_2 < \max(T_j) - b_1$. With the MVTFR data, $b_2 = 411.77$ and $\max(T_j) - b_1 = 692.97$, so that condition is met. Can $b_1 + b_2 > \max(T_j)$? In this analysis, $b_1 + b_2 = 649.31$. Now $\max(T_j)$ must always be at least as large as the average rank sum which is given by $T_{\text{avg}} = n(k+1)/2 = 19(52)/2 = 494$. So, if $494 \le \max(T_j) < 649.31$, then $b_2 > \max(T_j) - b_1$, and R_2 would not select more populations (states) than R_1. The same argument holds for the selection rules R_3 and R_4.

In the earlier study by Green and McDonald, using MVTFRs from 1982 through 2004, rule R_1 chose 13 states for the 'worst' subset. In this study, rule R_1 chose the same 13 states in addition to South Dakota (SD), Wyoming (WY), and Kentucky (KY). In the earlier study, rule R_3 chose 10 states for the 'best' subset. In this study, 9 of those 10 states were chosen (the exception being ND) and New York (NY), Vermont (VT), and California (CA) were added to this 'best' subset.

A PARAMETRIC SUBSET SELECTION OF STATES

In this section, a normal means parametric selection procedure will be used to contrast the inference with that of the nonparametric approach. This approach to subset selection was developed by Gupta.[20] With the additive model (1)

$$E(X_{ij}) = \mu + \theta_i + \beta_j. \tag{8}$$

Letting $\overline{X}_i = (\sum_j X_{ij})/n$, then $E(\overline{X}_i) = \mu + \theta_i + (\sum_j \beta_j)/n$. Since the quantity $\mu + (\sum_j \beta_j)/n$ is constant for all i, inference on the ordered θ_i can be efficiently based on the ordering of the means, $\overline{X}_i$.

The additive model (1) will be used with X_{ij} replaced with $f(X_{ij}) = X_{ij}^{0.3}$ based on the results given in Table 2. (The optimal Box-Cox λ is 0.37). Here, the ε_{ij} are assumed independent identically distributed normal variates with mean 0 and standard deviation σ. Residual displays from a two-way additive analysis of variance (ANOVA) are given in Figure 3. The residuals are symmetrically distributed with some outliers on the lower and upper ends. The 'raw' data now will be the fatality rates raised to the 0.3 power. Since our interest is selection of 'best' and 'worst' subsets, we will retain the 'outliers' and continue with a normal means selection process using the selection rule R_5 for the 'worst' population subset and R_6 for the 'best' population defined as follows:

R_5: Select the ith state iff $\overline{X}_i \ge \overline{X}_{[k]} - d$, $d > 0$

R_6: Select the ith state iff $\overline{X}_i \le \overline{X}_{[1]} + c$, $c > 0$.

The $\overline{X}_i$'s are the respective sample means of the 'raw' data and the $\overline{X}_{[i]}$'s are the ordered sample means. The positive constants d and c are chosen so that the $P(CS) \ge P^*$ for any configuration of the population (state) parameters, θ_i's. It can be shown that for a fixed P^*, d = c, and

$$d = h\sigma(2/n)^{1/2}, \tag{9}$$

where h is defined by the integral equation (7).

For $k = 51$, $n = 19$, and $P^* = 0.90$, the constants $d = c = 0.8436\sigma$. The value of σ is chosen to be 0.033 based on the two-way additive ANOVA of the transformed MVTFRs (i.e., the square root of the Mean Square for Error) as shown in Table 3. Then $d = c = 0.8436 \cdot (0.033) = 0.027839$. The means of the transformed rates are given in Appendix C. The maximum sample mean is 1.27436 (MS) and the minimum sample mean is 0.93089 (MA).

For selecting the 'worst' subset,

R_5: Select the ith state iff $\overline{X}_i \ge \overline{X}_{[k]} - d = 1.27436 - 0.027839 = 1.24652$.

The three states South Carolina (SC), Montana (MT), and Mississippi (MS) are chosen

For selecting the 'best' subset,

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

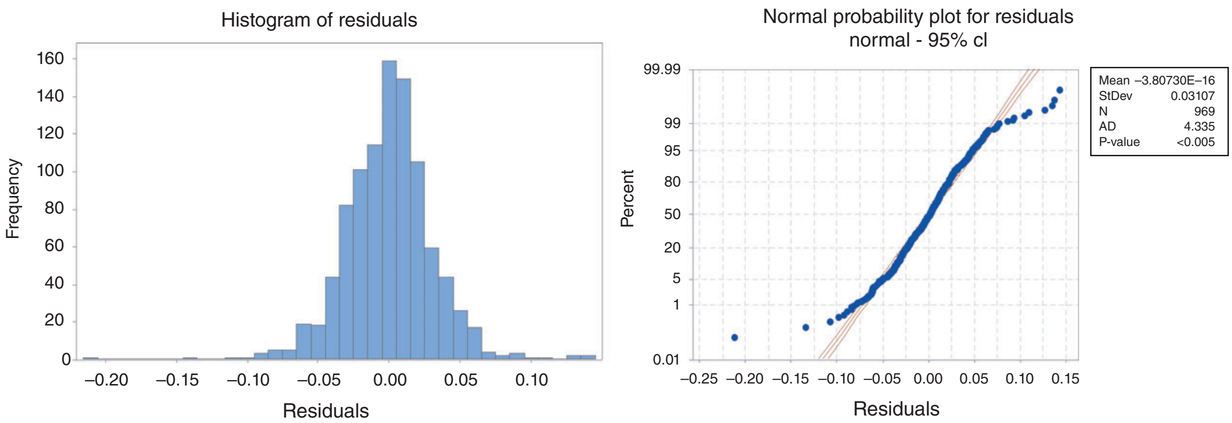


FIGURE 3 | Residual plots for $(\text{MVTFRs})^{0.3}$ from a two-way additive ANOVA. ANOVA, analysis of variance; MVTFRs, motor vehicle traffic fatality rates.

R_6: Select the ith state iff $\overline{X}_i \leq \overline{X}_{[1]} + c = 0.93089 + 0.027839 = 0.958729$.

Only the state of MA is chosen for the selected subset.

An advantage of the parametric approach over the nonparametric approach is that the parametric analysis explicitly utilizes the magnitudes of the data rather than simply their rank values. Thus, in this analysis, the normal means parametric approach results in a dramatic reduction in the number of states chosen for the selected subsets. If value of P^* is increased to 0.99, the chosen 'worst' subset would increase to five states, adding Louisiana (LA) and Arkansas (AR); and the chosen 'best' subset would remain at one state, MA.

A closer examination of the residual plots in Figure 3 (and listing of residuals not given here) shows that the smallest two residuals belong to District of Columbia (DC) and ND for the years 2012 and 1995, respectively. Furthermore, four of the smallest six residuals belong to DC and ND. At the other extreme, the largest three residuals belong to ND (2012, 2009, and 2011) and the next three largest belong to DC (2003, 2001, and 1994). Deleting DC and ND and following the same parametric analysis as in this section, with $k = 49$, $n = 19$, $h = 2.582$ (for $P^* = 0.90$), $\sigma = 0.033$, $d = c = 0.02764$, the exact same selection of states is made as with the inclusion of DC and ND. The residual plots of the residuals appear more normal-like with the two states deleted as shown in Appendix D. This examination of residuals and subsequent reanalysis of the data is simply one form of sensitivity analysis. In this case, the parametric selection of states is not affected by the deletion of those two states (DC and ND) which might be deemed outliers.

A BAYESIAN APPROACH TO THE SELECTION PROBLEM

In this section, a Bayesian approach is adopted and the population means are assumed to be stochastic. The idea is quite straightforward. A posterior distribution on the population means is used to simulate a large number of random draws, or realizations, of those means. With those draws, ordering probabilities of the population means can be estimated. And from these estimates, simple calculations can provide estimates of, e.g., the probability that a specific population mean is greater than all the other population means. There are

TABLE 3 | Two-Way ANOVA Table for the Transformed MVTFRs

Source	Degrees-of-Freedom	Sums of Squares	Mean Squares	*F* Ratios	*P* Values
State	50	6.41456	0.12829	123.55	0.000
Year	18	2.28450	0.12692	122.22	0.000
Error	900	0.93455	0.00104		
Total	968	9.63361			

ANOVA, analysis of variance; MVTFRs, motor vehicle traffic fatality rates.

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

many choices that can be made for the posterior distributions. One such approach, utilizing flat (or noninformative) prior distributions on the population means is illustrated here.

As shown in Gill[21] and many other Bayesian texts, the posterior distribution of the mean of the *i*th population (state), μ_i, is normal with mean $\overline{X}_i$ and standard deviation $\sigma/\sqrt{n}$, i = 1,…,k. In this situation, the Bayesian and frequentist results (via Central Limit Theorem) are very similar in form.

The relevant calculations become

$$\mathrm{P}(\mu_{\mathrm{m1}} \le \mu_{\mathrm{m2}} \le \ldots \le \mu_{\mathrm{mk}}), \tag{10}$$

where (m1,m2,…,mk) is any of the k factorial permutations of the integers (1,2,…,k). For example, for k = 4, there would be 4! = 24 such probabilities to calculate. This can be easily handled with WinBugs (an MCMC simulator) or R. *For frequentists, these calculations are meaningless*. This approach is used in both of the following subsections. In section *Example with k = 4*, all of the permutation probabilities (10) can be estimated with simulated draws of the posterior mean as k is small. In section *Bayesian Analysis of MVTFR*$^{0.3}$ *Using R*, with large k, applicable to the analysis of MVTFRs, a convenient function in R is used to identify which population (state) realizes the largest and smallest posterior mean on each simulation pass.

Example with k = 4

Suppose we have k = 4 populations with three observations from each of the populations yielding sample means of 2, 3, 4, and 5. Assume a common known standard deviation equal to 1 and a flat (noninformative) prior distribution on the population means. Using, for example, WinBugs, all 24 values of the probabilities given in Eq. (10) can be computed. By appropriate summing, the estimated values of $P[\mu_{\mathrm{i}} = \max(\mu_j)]$, i = 1, 2, 3, 4, are obtained. Table 4 gives the results of such computations for 6 of the 24 parameter permutations. The tabled values were generated with WinBugs using the model code given in Appendix E and specifying a large number (10^5) draws on the posterior means. The probabilities of the permutations, $P(\mu_{\mathrm{m1}} \le \mu_{\mathrm{m2}} \le \mu_{\mathrm{m3}} \le \mu_{\mathrm{m4}})$, are denoted by Pm1.m2.m3.m4 in Table 4.

TABLE 4 | A Sampling of WinBugs Estimates for Selection from Four Populations

P1.2.3.4 = 0.6844
P1.3.2.4 = 0.1031
P2.1.3.4 = 0.09304
P2.3.1.4 = 0.00274
P3.1.2.4 = 0.00278
P3.2.1.4 = 0.0012

Given the probabilities in Table 4, it now follows that $P(\mu_4$ is max) = 0.6844 + 0.1031 + … + 0.0012 = 0.8873, i.e., the sum of the six probabilities in the Table. In a similar manner, the calculations yielded $P(\mu_1$ is max) = 0.00006, $P(\mu_2$ is max) = 0.00364, and $P(\mu_3$ is max) = 0.10900. A complete probability distribution over all possible orderings of the population means is realized. This approach of calculating all the permutation probabilities is, from a practical vantage, limited to small values of k (say k ≤ 5 or 6). In our application to traffic fatality rates where k = 51, another Bayesian approach is more useful as described in the next subsection.

Bayesian Analysis of MVTFR$^{0.3}$ Using R

The power transformation makes plausible the negligible interaction assumption for the additive model. The 'state effects,' assuming flat normal priors, have a normal distribution centered at $\overline{X}_i$ and standard deviation $\sigma/\sqrt{n}$, i = 1,…,k. We now simulate in R a draw from each state, rank the results (using 'which.max' and 'which.min'), and repeat a large number of times (e.g., 10^6) to obtain P(MA is best) = 1.000; and P(MS is worst) = 0.698, P(MT is worst) = 0.300, and P(the worst is any other than MS or MT) = 0.002. The R code for these calculations is given in Appendix F.

The results of the Bayesian analysis herein presented are in close agreement with the results given in the fifth section with the parametric selection procedure. This is as expected since the choice of a noninformative prior distribution results in an analysis based on the likelihood function as is the parametric selection procedure.

CONCLUDING REMARKS

The subset selection procedures, parametric or nonparametric, select a random number of populations to include in the subsets on which a confidence statement can be attached. Subset size is a random variable dependent on the observed data. Determining the constants required to implement the selection rules does require determination of the Least Favorable Configuration (LFC), i.e., the configuration of population parameters which minimize the

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

probability of a CS. With two of the procedures used in this article, R_1 and R_3, that determination has been made only in the situation where the underlying parameter space is a 'slippage' space, i.e., all population parameters equal with the possible exception of one. However, there have been limited simulation studies, e.g., see Lorenzen and McDonald,[8] suggesting that the inference is valid in much more general settings. For all selection rules herein used, the LFC is that for which all population parameters are equal. The results for MVTFRs for the years 1994–2012 were compared to those of a similar analysis for the years 1982–2004.

The nonparametric selection rules choose a much larger subset than do the parametric procedures. And the conclusions from the Bayesian analyses are qualitatively closely aligned with those from the parametric selection procedures. This is not surprising as the nonparametric approach uses the ranks of the data, not the magnitudes. And as seen in Figure 3, there are outliers on the lower and upper end of the residual probability plot.

Bayesian procedures can yield a complete probability distribution over all orderings of the population parameters (e.g., means). There is a curse of dimensionality--k! gets large very quickly. However, using simulation capability in WinBugs and R, it is straightforward to generate a probability distribution over the populations as to which has the maximum (minimum) parameter. This was illustrated with MVTFRs from k = 51 states.

With respect to analyses of MVTFRs, there is an important research and application literature dealing with the identification of 'black spots,' or hazardous locations, to which safety measures can be applied to improve traffic safety. Hauer[22] and Montella[23] review statistical procedures for identification of such road sections or intersections. A Bayesian approach to investigate and evaluate ranking criteria for black spot identification is given by Lan and Persaud.[24] Bayesian multiple testing procedures for hotspot identification are given by Miranda-Moreno et al.[25] They use a dataset of highway-railway grade crossings to illustrate procedures incorporating both the posterior distribution of accident frequency and the posterior distribution of ranks. Cheng and Washington[26] utilize five tests for conducting performance assessments of hot spot identification methods and conclude that, among the methods investigated, an empirical Bayes method is superior. While these important hot spot identification problems have not been addressed in this article, the application of these ranking and selection methods, or those more broadly described by Gibbons et al.[27] and Gupta and Panchapakesan,[28] might well be pursued for such applications.

FURTHER READING

Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton: CRC Press; 2013. Albert J. *Bayesian Computation with R*. 2nd ed. New York: Springer; 2009.

REFERENCES

1. Green J, McDonald GC. Nonparametric subset selection procedures: applications and properties. *Am J Math Manag Sci* 2009, 29:413–436.
2. McDonald GC. Nonparametric selection procedures applied to state traffic fatality rates. *Technometrics* 1979, 21:515–523.
3. McDonald GC. Some Multiple comparison selection procedures based on ranks. *Sankhya: Indian J Stat Ser A* 1972, 34:53–64.
4. McDonald GC. The distribution of some rank statistics with applications in block design selection problems. *Sankhya: Indian J Stat Ser A* 1973, 35:187–204.
5. McDonald GC. Characteristics of block design selection procedures and a counterexample. *Sankhya: Indian J Stat Ser B* 1985, 47:47–55.
6. Gupta SS, Nagel K, Panchapakesan S. On the order statistics from equally correlated normal random variables. *Biometrika* 1973, 60:403–413.
7. Sivak M. Road safety in the individual U. S. States: current status and recent changes, Report No. UMTRI-2014-20, The University of Michigan Transportation Research Institute, Ann Arbor, MI, 2014. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/108252/103020.pdf?sequence=1. (Accessed August, 2016)
8. Lorenzen TJ, McDonald GC. A nonparametric analysis of urban, rural, and interstate traffic fatality rates. In: Santner TJ, Tamhane AC, eds. *Design of Experiments—Ranking and Selection*. New York: Marcel Dekker; 1984, 143–178.

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

9. Green J, McDonald GC, Rao N. Using selection procedures to analyze state traffic fatality rates. *Am J Math Manag Sci* 2006, 26:387–416.
10. Neter J, Kutner MH, Wasserman W, Nachtsheim CJ. *Applied Linear Regression Models*. 3rd ed. New York: McGraw-Hill; 1996.
11. Kuehl RO. *Design of Experiments: Statistical Principles of Research Design and Analysis*. 2nd ed. New York: Duxbury Press; 1999.
12. Winer BJ. *Statistical Principles in Experimental Design*. 2nd ed. New York: McGraw-Hill; 1971.
13. Christensen R. *Analysis of Variance, Design and Regression*. Boca Raton, FL: Chapman & Hall/CRC; 1998.
14. Tukey JW. One degree-of-freedom for non-additivity. *Biometrics* 1949, 5:232–242.
15. Woolsey TD. Adjusted Death Rates and Other Indices of Mortality. In: Linder FE, Grove RD, eds. *Techniques of Vital Statistics Reprint of Chapter I-IV of Vital Statistics in the United States*. National Office of Vital Statistics, Washington D.C. 20402, 1959, 60-90. Stock number 1722-00236.
16. Duffy EA, Carroll RE. United States Metropolitan Mortality, 1959–1961, PHS Publication No. 999-AP-39, U.S. Public Health Service, National Center for Air Pollution Control, 1967.
17. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: The MIT Press; 1975.
18. Gibbons DI, McDonald GC. Sensitivity of an air pollution and health study to the choice of a mortality index. In: Ghosh S, ed. *Statistical Design & Analysis of Industrial Experiments*. New York: Marcel Dekker; 1990, 153–174.
19. North Dakota Department of Transportation. 2014 North Dakota Crash Summary, NDOT, Safety Division, Bismarck, ND, 2014. https://www.dot.nd.gov/divisions/safety/docs/crash-summary.pdf. (Accessed August, 2016)
20. Gupta SS. On some multiple decision (ranking and selection) rules. *Technometrics* 1965, 7:225–245.
21. Gill J. *Bayesian Methods: A Social and Behavioral Sciences Approach*. 3rd ed. Boca Raton, FL: Taylor & Francis Group; 2015.
22. Hauer E. Identification of Sites with Promise. *Transp Res Rec* 1996, 1542–09:54–60.
23. Montella A. A comparative analysis of hotspot identification methods. *Accid Anal Prev* 2010, 42: 571–581.
24. Lan B, Persaud B. Fully Bayesian approach to investigate and evaluate ranking criteria for black spot identification. *Transp Res Rec* 2011, 2237: 117–125.
25. Miranda-Moreno LF, Labbe A, Fu L. Bayesian multiple testing procedures for hotspot identification. *Accid Anal Prev* 2007, 39:1192–1201.
26. Cheng W, Washington S. New criteria for evaluating methods of identifying hot spots. *Transp Res Rec* 2008, 2083:76–85.
27. Gibbons J, Olkin I, Sobel M. *Selecting and Ordering Populations: A New Statistical Methodology*. New York: John Wiley & Sons; 1977. Republished in the Classics in Applied Mathematics Series, No. 26 (1999), Society for Industrial and Applied Mathematics, Philadelphia.
28. Gupta SS, Panchapakesan S. *Multiple Decision Procedures*. New York: John Wiley & Sons; 1979. Republished in the Classics in Applied Mathematics Series, No. 44 (2002), Society for Industrial and Applied Mathematics, Philadelphia.

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

APPENDIX A. Motor Vehicle Traffic Fatality Rates (Fatalities per 100 Million VMT)

State	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Alabama	2.21	2.2	2.23	2.23	1.94	2.03	1.76	1.75	1.8	1.71	1.95	1.92	1.99	1.81	1.63	1.38	1.34	1.38	1.33
Alaska	2.05	2.11	1.97	1.76	1.55	1.74	2.3	1.89	1.82	1.98	2.02	1.45	1.49	1.59	1.27	1.3	1.17	1.57	1.23
Arizona	2.33	2.61	2.36	2.19	2.17	2.18	2.11	2.12	2.18	2.07	2.01	1.97	2.07	1.7	1.52	1.31	1.27	1.39	1.37
Arkansas	2.44	2.37	2.21	2.35	2.2	2.07	2.24	2.08	2.13	2.09	2.22	2.05	2.01	1.96	1.81	1.8	1.7	1.67	1.65
California	1.56	1.52	1.43	1.32	1.2	1.19	1.22	1.27	1.27	1.31	1.25	1.32	1.29	1.22	1.05	0.95	0.84	0.88	0.88
Colorado	1.74	1.84	1.71	1.62	1.6	1.54	1.63	1.73	1.71	1.48	1.45	1.26	1.1	1.14	1.15	1.01	1.96	0.96	1.01
Connecticut	1.14	1.13	1.1	1.19	1.12	1.01	1.11	1.03	1.04	0.95	0.93	0.88	0.98	0.92	0.95	0.71	1.02	0.71	0.75
Delaware	1.59	1.61	1.51	1.79	1.4	1.18	1.49	1.58	1.4	1.57	1.44	1.4	1.57	1.23	1.35	1.28	1.13	1.1	1.24
District of Columbia	2	1.67	1.59	1.8	1.63	1.18	1.37	1.81	1.33	1.87	1.15	1.29	1.02	1.22	0.94	0.8	0.67	0.76	0.42
Florida	2.2	2.19	2.12	2.08	2.05	2.06	1.99	1.77	1.76	1.71	1.65	1.75	1.65	1.56	1.5	1.3	1.25	1.25	1.27
Georgia	1.72	1.74	1.76	1.69	1.63	1.52	1.47	1.53	1.41	1.47	1.44	1.52	1.49	1.46	1.37	1.18	1.12	1.13	1.11
Hawaii	1.54	1.64	1.84	1.65	1.5	1.21	1.55	1.61	1.34	1.43	1.46	1.39	1.58	1.33	1.04	1.09	1.13	0.99	1.25
Idaho	2.15	2.13	1.99	2.01	1.97	1.99	2.04	1.84	1.86	2.05	1.77	1.85	1.76	1.6	1.52	1.46	1.32	1.05	1.13
Illinois	1.68	1.68	1.53	1.41	1.38	1.42	1.38	1.37	1.35	1.36	1.24	1.27	1.17	1.16	0.98	0.86	0.88	0.89	091
Indiana	1.59	1.49	1.49	1.36	1.42	1.46	1.25	1.27	1.09	1.15	1.3	1.31	1.27	1.23	1.11	0.9	1	0.98	0.99
Iowa	1.86	2.03	1.73	1.67	1.55	1.68	1.51	1.49	1.31	1.42	1.23	1.45	1.4	1.43	1.34	1.19	1.24	1.15	1.16
Kansas	1.79	1.76	1.89	1.82	1.82	1.95	1.64	1.75	1.78	1.64	1.57	1.44	1.55	1.38	1.29	1.31	1.44	1.29	1.32
Kentucky	1.95	2.07	1.98	1.97	1.91	1.75	1.75	1.83	1.95	1.99	2.04	2.08	1.91	1.8	1.74	1.67	1.58	1.5	1.58
Louisiana	2.25	2.31	2.37	2.44	2.3	2.28	2.3	2.2	2.09	2.13	2.08	2.14	2.17	2.19	2.03	1.84	1.59	1.46	1.54
Maine	1.51	1.49	1.32	1.45	1.42	1.28	1.19	1.33	1.47	1.39	1.3	1.13	1.25	1.22	1.06	1.1	1.11	0.95	1.16
Maryland	1.47	1.5	1.32	1.31	1.25	1.2	1.17	1.27	1.23	1.19	1.16	1.09	1.16	1.09	1.07	0.99	0.88	0.86	0.89
Massachusetts	0.94	0.92	0.83	0.87	0.78	0.8	0.82	0.9	0.86	0.86	0.87	0.8	0.78	0.79	0.67	0.62	0.64	0.68	0.62
Michigan	1.67	1.79	1.67	1.58	1.46	1.44	1.41	1.34	1.28	1.27	1.12	1.09	1.04	1.04	0.96	0.9	0.97	0.94	0.99
Minnesota	1.49	1.35	1.3	1.22	1.31	1.22	1.19	1.06	1.2	1.18	1	0.98	0.87	0.89	0.78	0.74	0.73	0.65	0.69
Mississippi	2.77	2.94	2.65	2.73	2.77	2.66	2.67	2.18	2.43	2.33	2.28	2.32	2.2	2.04	1.79	1.73	1.61	1.62	1.51
Missouri	1.9	1.87	1.88	1.89	1.81	1.64	1.72	1.62	1.77	1.81	1.64	1.83	1.59	1.43	1.41	1.27	1.16	1.14	1.21
Montana	2.22	2.28	2.12	2.82	2.47	2.24	2.4	2.3	2.59	2.41	2.04	2.26	2.34	2.45	2.12	2.01	1.69	1.79	1.72
Nebraska	1.75	1.61	1.8	1.77	1.79	1.64	1.53	1.36	1.64	1.54	1.32	1.43	1.39	1.32	1.09	1.15	0.98	0.95	1.1
Nevada	2.26	2.24	2.18	2.13	2.19	2.01	1.83	1.72	2.12	1.91	1.95	2.06	1.97	1.68	1.56	1.19	1.16	1.02	1.07
New Hampshire	1.13	1.11	1.22	1.12	1.11	1.18	1.05	1.15	1.01	0.98	1.26	1.24	0.93	0.96	1.06	0.85	0.98	0.71	0.84

State	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
New Jersey	1.26	1.27	1.31	1.23	1.15	1.11	1.08	1.08	1.1	1.05	0.99	1.01	1.02	0.95	0.8	0.8	0.76	0.86	0.79
New Mexico	2.18	2.29	2.25	2.21	1.91	2.05	1.9	2	1.97	1.92	2.18	2.04	1.88	1.54	1.39	1.39	1.38	1.36	1.43
New York	1.49	1.46	1.34	1.37	1.23	1.26	1.13	1.2	1.15	1.11	1.08	1.03	1.03	0.97	0.92	0.87	0.92	0.92	0.91
North Carolina	1.99	1.9	1.89	l.81	1.87	1.71	1.74	1.67	1.7	1.66	1.64	1.53	1.53	1.62	1.4	1.28	1.29	1.19	1.23
North Dakota	1.39	1.13	1.26	1.47	1.25	1.64	1.19	1.45	1.32	1.41	1.32	1.62	1.41	1.42	1.33	1.72	1.27	1.62	1.69
Ohio	1.4	1.35	1.35	1.39	1.36	1.36	1.29	1.29	1.31	1.17	1.15	1.2	1.11	1.13	1.1	0.92	0.97	0.91	1
Oklahoma	1.93	1.74	1.96	2.02	1.8	1.74	1.5	1.57	1.62	1.47	1.67	1.71	1.57	1.61	1.55	1.57	1.4	1.47	1.48
Oregon	1.68	1.91	1.73	1.62	1.61	1.19	1.33	1.42	1.26	1.46	1.28	1.38	1.35	1.31	1.24	1.11	0.94	0.99	1.01
Pennsylvania	1.56	1.57	1.52	1.59	1.48	1.52	1.49	1.49	1.54	1.48	1.38	1.5	1.41	1.37	1.36	1.22	1.32	1.3	1.32
Rhode Island	0.89	1	0.97	1.06	0.93	1.06	0.96	1.01	1.03	1.24	0.98	1.05	0.98	0.8	0.79	1.01	0.81	0.84	0.82
South Carolina	2.27	2.28	2.34	2.18	2.34	2.41	2.34	2.27	2.23	2.01	2.11	2.21	2.08	2.11	1.86	1.82	1.65	1.7	1.76
South Dakota	2.02	2.06	2.24	1.86	2.04	1.82	2.05	2	2.12	2.38	2.24	2.22	2.08	1.62	1.35	1.48	1.58	1.23	1.46
Tennessee	2.23	2.24	2.12	2.02	1.94	2.01	1.99	1.85	1.73	1.73	1.89	1.79	1.82	1.7	1.5	1.4	1.47	1.32	1.42
Texas	1.79	1.76	2.02	1.77	1.74	1.67	1.72	1.73	1.73	1.71	1.6	1.5	1.48	1.42	1.48	1.35	1.29	1.29	1.43
Utah	1.9	1.73	1.64	1.79	1.65	1.63	1.65	1.24	134	1.29	1.2	1.12	1.11	1.11	1.06	0.93	0.95	0.93	0.82
Vermont	1.25	1.71	1.38	1.48	1.58	1.38	1.12	1.17	0.98	0.83	1.25	0.95	1.11	0.86	1	0.97	0.98	0.77	1.07
Virginia	1.38	1.29	1.23	1.4	1.29	1.19	1.24	1.27	1.18	1.23	1.17	1.18	1.19	1.25	1	0.94	0.9	0.94	0.96
Washington	1.35	1.33	1.44	1.32	1.27	1.21	1.18	1.21	1.2	1.09	1.02	1.17	1.12	1	0.94	0.87	0.8	0.8	0.78
West Virginia	2.08	2.16	1.97	2.08	1.9	2.08	2.14	1.91	2.19	1.96	2.02	1.82	1.96	2.1	1.82	1.82	1.64	1.78	1.76
Wisconsin	1.42	1.45	1.44	1.33	1.26	1.31	1.4	1.33	1.37	1.42	1.31	1.36	1.22	1.27	1.05	0.96	0.96	0.99	1.04
Wyoming	2.15	2.41	1.94	1.81	l.92	2.42	l.88	2.16	1.95	1.79	1.77	l.88	2.07	1.6	1.68	1.4	1.66	1.46	1.33

© 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

APPENDIX B. State Rank Sums and Subsets of States Chosen by Nonparametric Rules

$P^{*} = 0.90$		Worst Selection Rules		Best Selection Rules	
STATE	RANK SUM	R1	R2	R3	R4
Massachusetts	23			X	X
Connecticut	90.5			X	X
Rhode Island	93.5			X	X
New Jersey	105.5			X	X
Minnesota	124.5			X	X
New Hampshire	135.5			X	X
Washington	170			X	X
New York	180.5			X	X
Maryland	218			X	X
Vermont	227.5			X	X
Virginia	231.5			X	X
California	258.5			X	X
Ohio	261				X
Michigan	303.5				X
Illinois	305.5				X
Indiana	311				X
District of Columbia	323.5				X
Wisconsin	325.5				X
Maine	333.5				X
Utah	350				X
Oregon	398				X
Colorado	433		X		X
Hawaii	445.5		X		X
North Dakota	452.5		X		X
Delaware	457		X		X
Nebraska	462		X		X
Pennsylvania	489.5		X		X
Iowa	489.5		X		X
Georgia	500		X		X
Missouri	606.5		X		
Kansas	606.5		X		
Texas	609.5		X		
North Carolina	615.5		X		
Oklahoma	649.5		X		
Alaska	654.5		X		
Florida	701	X	X		
Idaho	715.5	X	X		
Nevada	722.5	X	X		
Tennessee	745.5	X	X		
Alabama	760	X	X		
Kentucky	769	X	X		
Wyoming	773.5	X	X		
New Mexico	775.5	X	X		
South Dakota	787	X	X		
Arizona	822.5	X	X		
West Virginia	823.5	X	X		
Arkansas	888	X	X		
Louisiana	899	X	X		
South Carolina	911	X	X		
Montana	929	X	X		
Mississippi	930.5	X	X		

APPENDIX C. Ordered Means of State MVTFR$^{0.3}$

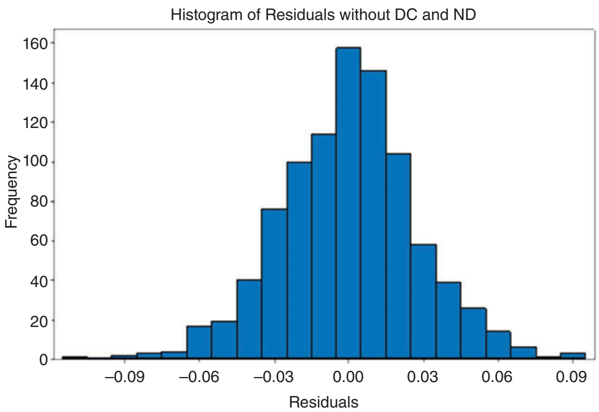
State	Mean	State	Mean
Massachusetts	0.93089	Pennsylvania	1.11572
Rhode Island	0.98628	Georgia	1.11814
Connecticut	0.99251	Iowa	1.11848
New Jersey	1.00680	Kansas	1.14954
Minnesota	1.00698	Missouri	1.15028
New Hampshire	1.01171	Texas	1.15057
Washington	1.02839	North Carolina	1.15157
New York	1.03323	Oklahoma	1.16115
Vermont	1.03750	Alaska	1.16758
Maryland	1.04387	Florida	1.17696
Virginia	1.04642	Idaho	1.18068
Ohio	1.05361	Nevada	1.18609
California	1.05533	Tennessee	1.18959
District of Columbia	1.06438	Alabama	1.19343
Indiana	1.06527	New Mexico	1.19984
Michigan	1.06677	Kentucky	1.20055
Illinois	1.06683	Wyoming	1.20078
Wisconsin	1.06903	South Dakota	1.20568
Maine	1.07259	Arizona	1.21516
Utah	1.07988	West Virginia	1.22244
Oregon	1.09229	Arkansas	1.23954
Colorado	1.10110	Louisiana	1.24497
Hawaii	1.10266	South Carolina	1.24829
Delaware	1.10750	Montana	1.26889
North Dakota	1.10844	Mississippi	1.27436
Nebraska	1.10871		

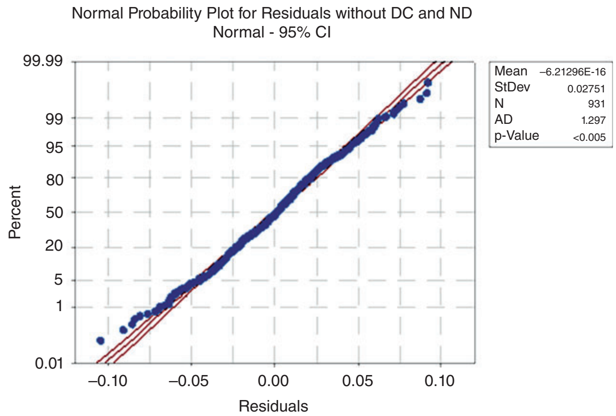
APPENDIX D. Analysis of MVTFR$^{0.3}$ with DC and ND Deleted

Table D1. Two-Way ANOVA Table for the Transformed MVTFRs with DC and ND Deleted

Source	Degrees-of-Freedom	Sums of Squares	Mean Squares	*F* Ratios	*P* Values
State	48	6.35409	0.13238	162.51	0.000
Year	18	2.21012	0.12278	150.73	0.000
Error	864	0.70380	0.00081		
Total	930	9.26801			

ANOVA, analysis of variance; DC, District of Columbia; MVTFR, motor vehicle traffic fatality rates; ND, North Dakota.





APPENDIX E. WinBugs Code for Calculations Related to Table 4

```
# Ranking & Selection for k = 4 Populations
    model {
    for (i in 1:3) {
        x1[i] ~ dnorm(m1,tau1)
        x2[i] ~ dnorm(m2,tau2)
        x3[i] ~ dnorm(m3,tau3)
        x4[i] ~ dnorm(m4,tau4)
    }
    m1 ~ dnorm(a,b)
    m2 ~ dnorm(a,b)
    m3 ~ dnorm(a,b)
    m4 ~ dnorm(a,b)
    tau1 < – pow(sigma1,-2)
    tau2 < – pow(sigma2,-2)
    tau3 < – pow(sigma3,-2)
    tau4 < – pow(sigma4,-2)
    p1.2.3.4 < – step(m2-m1)*step(m3-m2)*step(m4-m3)
    p1.2.4.3 < – step(m2-m1)*step(m4-m2)*step(m3-m4)
    p1.3.2.4 < – step(m3-m1)*step(m2-m3)*step(m4-m2)
    p1.3.4.2 < – step(m3-m1)*step(m4-m3)*step(m2-m4)
    p1.4.2.3 < – step(m4-m1)*step(m2-m4)*step(m3-m2)
    p1.4.3.2 < – step(m4-m1)*step(m3-m4)*step(m2-m3)
    p2.1.3.4 < – step(m1-m2)*step(m3-m1)*step(m4-m3)
```


© 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.

```
      p2.1.4.3 < − step(m1-m2)*step(m4-m1)*step(m3-m4)
      p2.3.1.4 < − step(m3-m2)*step(m1-m3)*step(m4-m1)
      p2.3.4.1 < − step(m3-m2)*step(m4-m3)*step(m1-m4)
      p2.4.1.3 < − step(m4-m2)*step(m1-m4)*step(m3-m1)
      p2.4.3.1 < − step(m4-m2)*step(m3-m4)*step(m1-m3)
      p3.1.2.4 < − step(m1-m3)*step(m2-m1)*step(m4-m2)
      p3.1.4.2 < − step(m1-m3)*step(m4-m1)*step(m2-m4)
      p3.2.1.4 < − step(m2-m3)*step(m1-m2)*step(m4-m1)
      p3.2.4.1 < − step(m2-m3)*step(m4-m2)*step(m1-m4)
      p3.4.1.2 < − step(m4-m3)*step(m1-m4)*step(m2-m1)
      p3.4.2.1 < − step(m4-m3)*step(m2-m4)*step(m1-m2)
      p4.1.2.3 < − step(m1-m4)*step(m2-m1)*step(m3-m2)
      p4.1.3.2 < − step(m1-m4)*step(m3-m1)*step(m2-m3)
      p4.2.1.3 < − step(m2-m4)*step(m1-m2)*step(m3-m1)
      p4.2.3.1 < − step(m2-m4)*step(m3-m2)*step(m1-m3)
      p4.3.1.2 < − step(m3-m4)*step(m1-m3)*step(m2-m1)
      p4.3.2.1 < − step(m3-m4)*step(m2-m3)*step(m1-m2)
      P[1] < − p1.2.3.4
      P[2] < − p1.2.4.3
      p[3] < − p1.3.2.4
      p[4] < − p1.3.4.2
      p[5] < − p1.4.2.3
      p[6] < − p1.4.3.2
      p[7] < − p2.1.3.4
      p[8] < − p2.1.4.3
      p[9] < − p2.3.1.4
      p[10] < − p2.3.4.1
      p[11] < − p2.4.1.3
      p[12] < − p2.4.3.1
      p[13] < − p3.1.2.4
      p[14] < − p3.1.4.2
      p[15] < − p3.2.1.4
      p[16] < − p3.2.4.1
      p[17] < − p3.4.1.2
      p[18] < − p3.4.2.1
      p[19] < − p4.1.2.3
      p[20] < − p4.1.3.2
      p[21] < − p4.2.1.3
      p[22] < − p4.2.3.1
      p[23] < − p4.3.1.2
      p[24] < − p4.3.2.1
      p.sum < − sum(p[])
}
List(a = 0,b = 0.001,x1 = c(1,2,3),x2 = c(2,3,4),x3 = c(3,4,5),x4 = c(4,5,6),
      sigma1 = 1,sigma2 = 1,sigma3 = 1,sigma4 = 1)
```

APPENDIX F. R Code for Bayesian Simulations Described in the Bayesian Analysis of MVTFR$^{0.3}$ Section

```
#R-code for Bayesian simulations of MVTFRs^0.3
# k = number of populations; n = number of simulations
#sigma = model sd; m = number of years
k = 51; n = 1000000;sigma = 0.0322241; m = 19
#sigma value is estimate from two-way ANOVA of rate^0.3
#mu values are means of (rate^0.3)
x < −c(rep(0,k))
y < −c(rep(0,n))
z < −c(rep(0,n))
err < −sigma/sqrt(m)
mu < −c(1.19343,1.16758,1.21516,1.23954,1.05533,1.10110,0.99251,1.10750,
1.06438,1.17696,1.11814,1.10266,1.18068,1.06683,1.06527,1.11848,
1.14954,1.20055,1.24497,1.07259,1.04387,0.93089,1.06677,1.00698,
1.27436,1.15028,1.26889,1.10871,1.18609,1.01171,1.00680,1.19984,
1.03323,1.15157,1.10844,1.05361,1.16115,1.09229,1.11572,0.98628,
1.24829,1.20568,1.18959,1.15057,1.07988,1.03750,1.04642,1.02839,
1.22244,1.06903,1.20078)
for (i in 1:n) {
  for (j in 1:k) {x[j] <−rnorm(1,mean = mu[j],sd = err)}
  y[i] <−which.min(x)
  z[i] <−which.max(x)
}
table(y)
table(z)
```

 © 2016 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals, Inc.