INVITED TALK - INFERENCE CONTROL PROBLEMS IN STATISTICAL DATABASE QUERY SYSTEMS

Lawrence H. Cox

Abstract: The advent of public use statistical database query systems raises problems of controlling inference of confidential information. Some of these problems are new while others present new challenges in terms of scalability of computational algorithms. We examine three problems: obtaining exact interval estimates of data withheld to address confidentiality concerns; confidentiality issues associated with the release of ordinary least squares regression models; and, confidentiality issues associated with the release of spatial statistical models based on ordinary kriging. For the first, we treat the database as one large multi-dimensional contingency table (large number of records, large dimension).

1. INTRODUCTION

National statistical offices (NSOs) collect, verify and refine statistical data to make reliable information available to policy makers and the public. By law or regulation and ethical practice, the NSO must preserve the *confidentiality* of data pertaining to individual entities such as persons, businesses, and health care providers.

Prior to 1960, NSOs made statistical information available primarily in the form of computed or estimated *tabulations,* defined by cross-classification of only one, two or a small number of variables. The NSO determined which tabulations to release, first in printed form and later also in electronic form. Confidentiality protection, more recently called *statistical disclosure limitation,* was accomplished by suppressing or combining selected tabulations or entire sets of tabulations or, less frequently, by altering tabulations slightly through rounding or incorporation of random noise. The NSO first determined which tabulations were worth

releasing and then released correspondingly less information in consideration of confidentiality and data quality concerns.

During the 1960s, first with the Continuous Work History Sample of the U.S. Social Security Administration, followed by Public Use Microdata Samples (PUMS) from the 1960 and subsequent U.S. Decennial Censuses, NSOs began releasing *statistical microdata files* comprising records pertaining to individual entities (mostly, persons). The data user was now free to create all conceivable summaries from the unit record data and, equally important, to fit statistical, demographic or econometric models to the microdata. Statistical disclosure limitation became focused on altering or removing selected microdata records. Longitudinal data presented confidentiality problems that remain largely unsolved. Emerging research is directed towards fitting the data to complex statistical models and releasing instead model-derived *synthetic microdata* and/or the models themselves. Disclosure limitation for tabulations and microdata are provably complex theoretically and computationally.

NSOs are considering allowing data users direct access to statistical databases, either on a public or restricted access basis, via a *statistical database query system.* This heightens confidentiality risk and will motivate disclosure limitation research in coming decades. In this paper, we investigate through examples some of the confidentiality and data useability problems raised by the advent of statistical database query systems. Several problems are illustrated by specialized examples. We focus on two query paradigms: tabulations from a database organized as a large multi-dimensional contingency table (Section 4) and simple statistical models derived from the database, namely, ordinary least squares regression models and best linear unbiased prediction *(kriging)* models for spatial data (Section 5). Section 6 contains concluding comments.

2. THE STATISTICAL DATABASE

For purposes here, a *statistical database* is equivalent to an *n-dimensional contingency table:* an enumeration of the units from a sample or population with respect to n cross-classified categorical variables. Each categorical variable i comprises d_i mutually exclusive and exhaustive *characteristics* c_{ik}. The *size* of the n-dimensional contingency table is $d_1 d_2 \cdots d_n$. Each *internal entry* $t_{i_1 i_2 \ldots i_n}$ of the table equals the number of units with characteristics $(i_1, i_2, \ldots, i_n)$. Internal entries therefore assume nonnegative integer values. This characterization is general and flexible. If every record in the underlying microdata file is uniquely identified by a combination of characteristics, then the characterization encompasses the

underlying microdata file. If not, at least in principle the same characterization is achieved by including an additional dimension defined by a unique identifier, such as social security number.

The table has many *marginal totals* corresponding to sums along one or more dimensions, *k-dimensional marginal totals* are totals along (n - k) dimensions. General mathematical notation for marginal totals is available, but somewhat cumbersome. Section 4 deals with complexities in n-dimensional tables, namely, properties that hold, e.g., in two dimensions, but fail entirely or in certain instances in higher dimensions. Examples are drawn from three and four dimensional tables and notation provided as needed.

3. CONFIDENTIALITY ISSUES IN STATISTICAL DATABASES

If a sample or population unit *(entity)* has one or more characteristics unique from those of the other units, then a third party potentially can identify the entity based on these *identifying characteristics.* In some instances, the simple act of identification is a breach of confidentiality. More typically, identification is based on fewer than the full set of n characteristics, resulting in disclosure of the remaining nonidentifying characteristics. If precisely two entities possess certain characteristics, then each potentially can identify the other and disclose confidential information. In general, *statistical disclosure* in contingency tables occurs when small counts are released or can be inferred. What constitutes small varies from one NSO to another. Traditional *threshold rules* are five (U.S. Census Bureau) and three (U.S. Internal Revenue Service and at Statistics New Zealand).

The number of entries in a n-dimensional contingency table typically is large and grows quickly with increasing dimension n. For example, even with all categorical variables dichotomous, the number of internal entries in a 30-dimensional table exceeds one billion. Most internal entries and higher dimensional marginal totals are likely to be small, in fact, zero or one. In this context, our notion of a *statistical database query system* is as follows. The database user can query the system as often as it likes, but each request must be for a marginal total. Of course, correct answers cannot be provided to queries corresponding to marginal totals not exceeding the threshold, but typically doing so in and of itself does not prevent a third party from deducing small entries, due to the additive structure of the table. Further *disclosure limitation* is required.

In two dimensional tables, it is possible to *round* all entries and totals in a manner that preserves additivity of internal entries to marginal totals. If all entries are rounded to multiples of the threshold, then disclosure limitation is complete. Similarly, it is possible to *perturb* entries slightly using additive random noise while preserving additivity. Small values remain, but the imprecision introduced through the perturbation is regarded as sufficient for disclosure limitation. Unfortunately, as demonstrated in the next section, it is not always possible to round or perturb entries in this manner in dimension n > 2. A third disclosure limitation method, *complementary suppression,* viz., the process of selectively suppressing entries to mask small entries, is complicated (indeed, *NP-hard*) even in two dimensions.

One approach to disclosure limitation in an n-dimensional statistical database is to answer only queries corresponding to lower dimensional marginal totals. The confidentiality issue is then whether the released totals can be used to infer small values. There are three aspects to this problem.

The first is: Can small values be inferred deterministically? This would be accomplished through manipulation of linear (additive) relationships between entries and the released marginal totals. This is essentially a problem in mathematical programming: Is the *feasible region* delimited *(constrained)* by the released marginals and nonnegativity of entries sufficient to ensure that each entry takes on at least one value at or above the threshold? Normally, this would correspond to a sequence of linear programming problemsBone to minimize and one to maximize each internal entry or marginal of interest over the feasible region, resulting in *exact bounds* $[\min\{t_{i_1,\ldots,i_n}\}, \max\{t_{i_1,\ldots,i_n}\}]$ for internal entries. This is a challenging but for the most part computationally tractable undertaking. Unfortunately, because entries must be integer, to yield *exact integer bounds* $[\min\{t_{i_1,\ldots,i_n} : t \text{ integer}\}, \max\{t_{i_1,\ldots,i_n} : t \text{ integer}\}]$, the NSO apparently is confronted with a massive integer programming problem, impossible to solve in general. This is illustrated by specialized examples and explored in Section 4.

The second aspect of the problem is: Can small values be inferred probabilistically? This would be accomplished using distributional models from the theory of log linear models and simulation. Some of the underlying mathematical issues here overlap with those raised in exact integer bounding. This problem is not addressed further here. The third aspect of the problem is: How to manage the query response strategy? The confidentiality problem is dynamic, namely, the response to successive new query potentially increases information about unreleased internal entries and marginals. One solution is to respond to queries on a flow basis, refusing any query that breaches confidentiality, and ending when no further queries can be answered safely. Another approach is to predetermine a (maximal)

set of queries that can be mutually answered safely and only to release information in response to these queries. Both approaches are computationally intensive and complex. These problems are worthy of investigation but not addressed further here.

4. PROPERTIES OF HIGH DIMENSIONAL TABLES

This section comprises a series of examples demonstrating the failure in higher dimensions of properties enjoyed by two-dimensional tables. Attempt is made to keep examples as uncomplicated as possible in order to emphasize essential features. All examples are of modest size and, with the exception of two four-dimensional table, are three-dimensional.

Cox and Ernst [2] demonstrate that in two-dimensional contingency tables *controlled rounding,* viz., rounding entries to a fixed integer rounding base while assuring that rounded and original entries differ by less than the base and that additivity to marginals is preserved, always can be accomplished. In addition, it is possible to ensure that any original entry equal to a multiple of the base remains fixed *(zero-restrictedness property)* [1]. Figure 1 depicts the internal entries of a three-dimensional table of size 2x2x2. Examination reveals that zero-restricted controlled rounding is not possible for Figure 1, and consequently is not assured in three and higher dimensions. Ernst [7] exploits this fact to construct a three-dimensional table for which a controlled rounding does not exist.

$$\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} 0 & 0.5 \\ 0.5 & 0 \end{pmatrix}$$

Figure 1. Zero-restricted controlled rounding fails in three dimensions

Controlled random perturbation is based on selecting a small positive perturbation value and alternately adding and subtracting it to/from original values while preserving additivity to marginals. Zero counts cannot be reduced, and therefore random perturbation must be zero-restricted. Arguments entirely analogous to those for controlled rounding show that controlled random perturbation is always possible in two-dimensional tables. Cox [4] demonstrates that controlled perturbation fails in three and higher dimensions. Consider Figure 2, a three-dimensional table of size 3x3x3. The * symbol denotes any positive value. It is not possible to alternate +/- movement of a positive quantity between nonzero values (*) while

preserving additivity to the table marginals. Controlled perturbation therefore fails.

$$\begin{pmatrix} 0 & * & * \\ * & 0 & * \\ * & * & 0 \end{pmatrix} \quad \begin{pmatrix} * & * & 0 \\ * & 0 & * \\ 0 & * & * \end{pmatrix} \quad \begin{pmatrix} * & 0 & * \\ 0 & 0 & 0 \\ * & 0 & * \end{pmatrix}$$

Figure 2. Controlled random perturbation fails in three-dimensions (* = positive entry)

Two vectors of nonnegative integers whose entries add to a common value are *consistent.* In two dimensions, a consistent pair of integer vectors assures the existence of one or more two-dimensional contingency table whose one-dimensional marginal (row and column) totals are given by the respective vectors. However, in n-dimensions, n consistent vectors of nonnegative integers do not necessarily comprise the (n-1)-dimensional marginal totals for any n-dimensional contingency table. Consider the three-dimensional table of Figure 3 (Vlach 1986) of size 3x4x6. Here, consistent integer two-dimensional marginals define a unique nonnegative table in which all entries are not integer. Consistent integer marginals can lead to an entirely *infeasible* situation, viz., no integer or continuous table exists; see Figure 4. In both examples, the + sign indicates the dimension over which the marginal is computed:

in Figure 3, $t_{+61} = \sum_{i=1}^{4} t_{i61} = 0$ and, in Figure 4, $t_{13+} = \sum_{k=1}^{3} t_{13k} = 3$.

$$(t_{ij+}) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad (t_{i+k}) = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}, \quad (t_{+jk}) = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

Figure 3. Consistent integer marginals fail to assure a feasible int three-dimensional table

$$(t_{ij+}) = \begin{pmatrix} 1 & 1 & 3 \\ 1 & 3 & 1 \\ 3 & 1 & 1 \end{pmatrix}, (t_{+jk}) = (t_{i+k}) = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}$$

Figure 4. Consistency fails to assure any feasible three-dimensional table

Assessment of disclosure risk in statistical tables and tabulations, referred to as *disclosure audit,* is the process by which to address the first question: Is the table safe from deterministic attempts to infer small values? This requires a mechanism for determining exact lower and upper bounds for each internal entry. In two dimensions, this is accomplished using simple formulae [3,4]. In higher dimensions, such formulae are not available except in specialized cases. It might appear that exact bounds could be computed using linear programming: For each internal entry t, solve one linear program to compute min $\{t\}$ and a second to compute max $\{t\}$. This is tractable computationally and can be accomplished with far fewer optimizations if interrelationships between bounds are exploited. This process would be sufficient for disclosure audit under any of the following three conditions.

One, if all *extremal points* of the linear programming polytope were integer-valued. Two, if every exact lower and upper bound occurred at one or more integer-valued points of the polytope, and an algorithm available to direct the linear program to one such point for each bound. Three, the *integer rounding property* (IRP) (Nemhauser and Wolsey 1988, 594-598) holds for each bound, viz., the exact integer bound corresponds to rounding the exact continuous bound down or up, respectively, to the nearest integer. The first condition holds in two dimensions, and therefore so do the second and third.

Unfortunately, all three conditions fail in higher dimensions, meaning that linear programming is not a viable method on which to base procedures for disclosure audit in general higher dimensional tables.

Failure of the first condition is illustrated in Figure 5, which displays all prescribed two-dimensional marginals for a set of 4x4x4 three-dimensional tables. Failure of the second condition is illustrated by Figure 6, which displays a noninteger extremal solution at which $\max\{t_{312} : t \text{ integer}\} = 1$

is achieved on the polytope of 3x3x3 three-dimensional tables with all one-dimensional marginals prescribed.

$$t_{ij+} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad t_{i+k} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad t_{+jk} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

Figure 5. Table with fractional continuous exact bound ($\max\{t_{331}\} = 1/2$)

$$t_{ij1} = \begin{pmatrix} 0 & 0 & 1.5 \\ 0 & 0 & 0 \\ 0.5 & 0 & 0 \end{pmatrix}, \quad t_{ij2} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad t_{ij3} = \begin{pmatrix} 0 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0 \end{pmatrix}$$

Figure 6. Noninteger $\text{intmax}\, t_{312}$; marginals $t_{i++} = t_{j++} = t_{++k} = (2 \quad 1 \quad 2)$

Failure of the integer rounding property is illustrated by several instructive examples. Figure 7 has a unique integer solution for which $t_{321} = 1$. However, the continuous minimum of this entry equals zero, and the integer rounding property fails. Figure 7 can be viewed as a table with suppressions, viz., original unsuppressed entries were subtracted from marginal entries and replaced by zeroes. Examples involving zero-restrictions are instructive in examining tables with suppressions, but zero-restrictions are not necessary to demonstrate failure of the integer rounding property. Figure 8 displays internal entries for a 2x2x2x2 table (Sturmfels 2002). This solution is the unique totally integer solution satisfying the corresponding two-dimensional marginal totals, despite the fact that these marginals define a feasible region in 16-dimensional space formed by intersection of a five-plane with the first orthant. The integer rounding property fails because

$$\max\{t_{1121}\} = 5/3 \text{ but } \max\{t_{1121} : t \text{ integer}\} = 0.$$

The continuous optimum in Figure 8 exceeds the integer optimum by more than one unit. This raises the question as to whether the continuous and integer maximum (or minimum) (the *integer programming gap*) can be arbitrarily far apart. This is important because, the farther apart they are, the less information about integer optima are contained in the continuous optima obtained via linear programming. A related question, posed by Figures 5 and 6, deals with the frequency of fractional optima. Further empirical evidence is provided in simulation experiments of Fagan [8] which revealed a 4x4x4x4 table with suppressions (too complex to represent here) for which several entries have integer minimum equal to zero, but continuous minima equal to 8/3, with many fractional optima, and for which the integer rounding property fails a total of 120 out of a possible 350 times. Also of interest is that, whereas linear programs achieve all values in the feasible range for an entry, is this also the case for the integer feasible range? Recent theoretical work has shown that the integer programming gap can be large [9] and furthermore that gaps can exist within the sequence of feasible integer values achieved by any particular table entry [6].

$$t_{i+k} = t_{ij+} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad t_{+jk} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix},$$

$$t_{211} = t_{222} = t_{231} = 0$$

Figure 7. IRP fails with zero-restriction: unique int. sol. $t_{321} = 1$ but $\min\{t_{321}\} = 0$

$$t_{ij11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \; t_{ij12} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \; t_{ij21} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; t_{ij22} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

Figure 8. Unique 4-D int. sol., fixed 2-D marginals: IRP fails: $\max\{t_{1121}\} = 5/3$

5. LINEAR AND SPATIAL PREDICTION USING STATISTICAL DATABASES

5.1 Ordinary least squares regression

An alternative output model for a statistical database is to release only regression coefficients as requested by users. Refusing, perhaps, to release regressions representing nearly perfect fit, this appears to be a safe release strategy. While for the most part this may be so, it is possible to construct scenarios under which disclosure occurs. Such scenarios, while unlikely to occur in practice, are instructive towards developing strategies for safe release. One such scenario is presented in the next paragraph.

Under simple linear ordinary least squares regression, assume that the user has requested regression of Y (say, income) on X (say, age) for all p database units with specific characteristics (say, statisticians in a particular city under the age of 80). The database returns a no-intercept model with regression coefficient β. Next, the user requests the same regression, but this time for all (p + m) database units satisfying more general characteristics (say, statisticians in the city under the age of 90). The database returns regression coefficient β^*. x, y denote the X- and Y-means of the m additional database records. Then,

$$\overline{Y}_{90} = \overline{X}_{90}\beta^*$$
$$(p\overline{Y}_{80} + m\overline{y})/(p+m) = \overline{X}_{90}\beta^*$$

Thus,

$$\overline{y} = ((p+m)/m)\overline{X}_{90}\beta^* - (p/m)\overline{Y}_{80}$$
$$= \overline{x}\beta^* + (p/m)(\overline{X}_{80}\beta^* - \overline{Y}_{80})$$

viz., $\overline{y}$ can be precisely determined. If m = 1 and the one statistician in the city of age 80-90 can be identified, then that statistician=s income is precisely determined. If m = 2, then either of the two elderly statisticians could subtract his or her income from y and again precisely determine the income of the other statistician. In general, if m is small, some disclosure is possible.

The question arises: Does adding noise to the x-variables limit disclosure in regression outputs? The simple linear regression is: $\mathbf{Y} = \mathbf{X}\beta$. Add zero-mean IID noise to the X-data $\mathbf{X_p} = \{\mathbf{x}_i\}_{i=1}^{p}$. In lieu of releasing the true

regression, the NSO generates zero-mean IID noise ε_p and creates p noisy data points $(x_k + \varepsilon_k, y_k)$. Simple linear regression on the noisy data results in the regression model:

$$\begin{aligned} Y = (X + \varepsilon)\gamma &= X[\mathrm{Cov}(X_p + \varepsilon_p, Y_p)/\mathrm{Var}(X_p + \varepsilon_p)] \\ &= X(1/(1 + \mathrm{Var}(\varepsilon_p)/\mathrm{Var}(X_p))\beta \end{aligned}$$

The user now requests an updated regression that in addition includes m additional data points:

$$\{ x_{p+k} \}_{k=1}^{m}$$

m additional noisy data points $\{ (x_{p+k} + \varepsilon_{p+k}, y_{p+k}) \}_{k=1}^{m}$ are created and an updated regression performed:

$$Y = (X + \varepsilon_{p+m})\gamma^* = X(1/(1 + \mathrm{Var}(\varepsilon_{p+m})/\mathrm{Var}(X_{p+m}))\beta^*$$

Often $Var(\varepsilon_{p+m})$ is known, and disclosure can be achieved as in the first section. Otherwise, as $Var(\varepsilon_{p+m})/\mathrm{Var}(X_{p+m})$ is small, approximate disclosure is possible.

5.2 Spatial statistical models based on ordinary kriging

Ordinary kriging is a method for best linear unbiased prediction of spatially referenced data. Observations $\{Z(x_1), ..., Z(x_s)\}$ are made at known locations $\mathbf{X} = \{x_1, ..., x_s\}$ identified , e.g., by latitude and longitude, and are fit to a covariance model $Cov(Z(x), Z(x'))$, from which a spatial *(kriging)* model is developed and used to predict the value of Z(x) at unobserved locations x. See [5] Chapter 3 for details. If, e.g., Z is Gaussian, then the best linear unbiased predictor is given by:

$$\hat{Z}(x) = \sum_{i=1}^{s} \lambda_i(x) Z(x_i) \text{ with}$$

$$\sum_{k=1}^{s} Cov(Z(x_i), Z(x_k))\, \lambda_k(x) = Cov(Z(x), Z(x_i)) \text{ and } \sum_{i=1}^{s} \lambda_i(x) = 1.$$

The confidentiality issue is whether it is safe for the NSO to release the kriging model. The answer is no: Because $\hat{Z}(x_i) = Z(x_i)$ and because locations are typically public knowledge, release of the kriging model results in exact disclosure of Z-data at the observed locations **X**.

What disclosure limitation options are available to the NSO? It is not possible to add noise ε to the locations, as the $Z(x_i + \varepsilon_i)$ are unknown. One possibility is as follows: (1) Krige based on $\{Z(x_i)\}$, resulting in $\hat{Z}(x)$. (2) Generate zero-mean IID noise ε. (3) Krige based on $\{Z(x_i + \varepsilon_i)\}$, resulting in $\hat{Z}_\varepsilon(x)$. (4) Release $\hat{Z}_\varepsilon(x)$.

A second, possibility is: (1) *Jiggle* the covariance matrix, viz., given $\mathbf{K} = (\mathrm{Cov}(Z(x_i), Z(x_j))$, create $\mathbf{K}_\varepsilon = \mathbf{K} + \varepsilon$. (2) Krige based on $\mathbf{K}_\varepsilon$.

However, this is tricky as the effects of small perturbations to entries of **K** on covariance and the resulting spatial model are unclear, viz., it is not clear if or how to ensure that $|Z(x_i) - \hat{Z}_\varepsilon(x_i)|$ is sufficiently large, but not too large.

6. CONCLUDING COMMENTS

It can be argued that the next evolution in the release of statistical data by NSOs is statistical database query systems. This moves the NSO into the arena of releasing tabulations from high dimensional and linked tabular structures. This on the one hand magnifies disclosure risk and on the other based on evidence presented here presents potentially significant theoretical and computational challenges to the NSO as it attempts to assess and control user inference of confidential information.

Strategies for releasing statistical models in lieu of original data or tabulations have been proposed to address confidentiality concerns. Based on evidence gained by examining linear regression and spatial prediction models, we conclude that the advantages and limitations of doing so need to be carefully assessed. However, as demonstrated here, new and potential inference control strategies are worth pursuing.

References

[1] Causey, B.D., Cox, L.H. and Ernst, L.R. Applications of transportation theory to statistical problems, *J. Amer. Stat. Assoc.* 80: 903-909, 1985.

[2] Cox, L.H. and Ernst, L.R. Controlled rounding, *INFOR* 20: 423-432, 1982

[3] Cox, L.H. Bounds on entries in 3-dimensional contingency tables subject to given marginal totals, in *Inference Control in Statistical Databases, Lecture Notes in Computer Science 2316,* J. Domingo-Ferrer, ed., Springer-Verlag, Heidelberg, pp. 21-33, 2002.

[4] Cox, L.H. Properties of multi-dimensional statistical tables, *J. Stat. Plan. and Inf.* 117: 251-273, 2003.

[5] Cressie, N.A.C. *Statistics for Spatial Data,* Wiley-Interscience, New York, 1993.

[6] De Loera, J. and Onn, S. All rational polytopes are transportation polytopes and all polytopal integer sets are contingency tables, in *Proceedings of the 10th Mathematical Programming Society Symposium on Integer Programming and Combinatorial Optimization, Lecture Notes in Computer Science,* Springer-Verlag, Heidelberg, 2004 (to appear).

[7] Ernst, L.R. Further applications of linear programming to sampling problems, Technical ReportBCensus/SRD/RR-89-05, Washington, DC, U.S. Census Bureau, Department of Commerce, 33 pp. ,http://www.census.gov/srd/www/byname.html, 1989

[8] Fagan, J.T. Personal communication, July 16, 2002.

[9] Hosten, S. and Sturmfels, B. Computing the integer programming gap, Manuscript, 23 Jan. 03, 17pp., rXiv.math.OC/0301266, 2003.