Statistical Methods in Medical Research 2005; **14**: 3–16

On selection and composition in small area and mapping problems

Nicholas T Longford SNTL, Leicester, UK

Maps in which small areas, such as districts, are represented by colours, shades or symbols with sizes determined by the values of estimates are regarded as an indispensable graphical output of analyses concerned with the geographical detail of economic, social, ecological and epidemiological phenomena. The distortion of the distribution of the district specific quantities in such maps, due to misrepresentation of the uncertainty about the estimated values, is discussed, and an alternative based on drawing so-called plausible maps is described. We highlight the pervasive and nonignorable nature of the selection process that identifies the quantity (target) to be estimated. A problem specific to disease mapping is what action, often one of a discrete set, to take in response to the results of an analysis. We argue that the costs (values) associated with correct and incorrect decisions should be integrated in the analysis and, when an analytical treatment is not feasible, plausible scenarios played out by simulations.

1 Introduction

Disease mapping and small area estimation are two successful applications of composite estimators. In their general forms, they can be interpreted as (linear or convex) combinations of estimators that would be suitable in some extreme settings. For example, in empirical Bayes methods, the direct estimator, based on the data only for the area concerned, is combined with the national estimator that is based on all the data. The direct estimator is unbiased but has a large variance, whereas the national estimator is biased for the area, but has a much smaller sampling variance.

In contrast, much of statistical practice is firmly wedded to selection (of models, estimators, and the like), motivated by hypothesis testing and related model selection procedures. Section 2 discusses this dichotomy in detail and reinterprets some estimators used in disease mapping as composite estimators. The purpose of the discussion is to show that complex modelling can be avoided if the estimators it yields are described as compositions. We pursue a general argument that all substantial sources of uncertainty have to be represented in modelling and accounted for in inferential statements. These sources include uncertainty about the estimates given the selected model, as well as about models themselves. Section 3 develops this theme further by considering the selection of the target (the quantity to be estimated) as a distinctly nonignorable process; that is, the way we select the target has a profound influence on the distribution of the estimator. Section 4 applies these principles to inferences about extremes and searches for evidence of an external agent that influences the outcomes. Section 5 deals

Address for correspondence: Nicholas T Longford, SNTL, 23 Fairstone Hill, Oadby, Leicester LE2 5RL, UK. E-mail: ntl@sntl.co.uk

10.1191/0962280205sm385oa

with making decisions about intervention based on incomplete information (uncertainty) and argues for integration of the costs in the statistical analysis.

Some of the algebra in Section 2 is presented for completeness, and can be skipped. The principal nontechnical message of the section is that the traditional model selection is not conducive to efficient estimation and unbiased (honest) assessment of the precision of the estimators. The alternative proposed combines model based (or any other) estimators of the target quantity. Sections 3–5 do not depend on Section 2 directly, as they are applicable to any setting that involves inferences about collections of unknown quantities. However, they discuss related undesirable features of selection of targets and emphasize the need for accounting for all sources of inferential uncertainty. When a collection of targets is considered, selection among them is a nontrivial source.

2 Selection and composition

Estimation of a population quantity θ_d in each of a set of geographical units (districts) $d = 1, \ldots, D$ of a domain (country) is a problem common to applications concerned with the spatial (geographical) distribution of a phenomenon, such as unemployment, crime, a medical condition or natural events (earth tremors or storm damage). In these examples, there is an obvious domain counterpart of θ_d; for instance, the national prevalence of a medical condition. It is denoted by θ.

Two trivial approaches to estimating district level quantities θ_d are *direct estimation* and *pooling*. In direct estimation, the only information used for estimating θ_d is the data from district d, whereas by pooling, θ_d is estimated for each district by an estimate for its domain counterpart, $\hat{\theta}$. The direct estimator is (usually) unbiased but, involving little data, it has a large sampling variance. In contrast, the pooled estimator has much smaller sampling variance but, involving data from outside the district, is biased. The bias is specific to the district. Attempts to decide which estimator to apply, $\hat{\theta}_d$ or $\hat{\theta}$, and for which districts, are not very effective because they can, at best, match the more efficient of the two estimators. Such a *selected* estimator is formally defined as

$$\hat{\theta}_d^{\dagger} = (1 - I_d)\hat{\theta}_d + I_d\hat{\theta}\,,$$

where I_d is the indicator of the selection ($I_d = 1$ when $\hat{\theta}$ is selected and $I_d = 0$ otherwise); $\hat{\theta}_d^{\dagger}$ is a *mixture* of the *constituent estimators* $\hat{\theta}$ and $\hat{\theta}_d$. This formulation of $\hat{\theta}_d^{\dagger}$ makes the (estimator or model) selection process explicit, and implies that it has an impact on the properties of the resulting estimator. The properties are difficult to establish in general, because I_d, a random variable, is usually correlated with both $\hat{\theta}$ and $\hat{\theta}_d$.

When I_d is independent of both $\hat{\theta}$ and $\hat{\theta}_d$, for instance, when the choice made is not informed by the data, elementary operations yield the identity

$$\mathrm{MSE}(\hat{\theta}_d^{\dagger}; \theta_d) = (1 - p_d)\mathrm{MSE}(\hat{\theta}_d; \theta_d) + p_d\mathrm{MSE}(\hat{\theta}; \theta_d)\,, \tag{1}$$

where $p_d = P(I_d = 1)$. We include the target as an argument of MSE because a statistic can be used as an estimator of several targets; in particular, $\mathrm{MSE}(\hat{\theta}; \theta_d) \neq \mathrm{MSE}(\hat{\theta}; \theta)$, unless $\theta_d = \theta$. The identity in Equation (1) implies that when selection is independent of the constituent estimators, the selected estimator cannot be more efficient than both constituent estimators. By ignoring the uncertainty associated with the selection of the estimator we falsely claim to have estimated the target $\hat{\theta}_d$ by the more efficient of the candidates $\hat{\theta}_d$ and $\hat{\theta}$. More precisely, we would claim that the MSE of $\hat{\theta}^{\dagger}$ is estimated without bias by

$$\hat{s}^2_{\dagger,d} = (1 - I_d)\hat{s}^2_d + I_d\hat{s}^2,$$

where $\hat{s}^2$ and $\hat{s}^2_d$ are the estimators of the MSEs (or sampling variances) of θ_d assuming that the districts have identical values of θ_d ($\theta_1 = \theta_2 = \ldots = \theta_D$) or not, respectively. Note that $\mathrm{MSE}(\hat{\theta};\, \theta_d) > s^2$ because $\hat{\theta}$ is biased for θ_d, so $\hat{s}^2$ is biased for $\mathrm{MSE}(\hat{\theta};\, \theta_d)$. Usually $\hat{s}^2_{\dagger,d}$ underestimates $\mathrm{MSE}(\hat{\theta}^{\dagger}_d;\, \theta_{\dagger})$. This should be interpreted as dishonesty, because the inferential statement claims more than what is justified. Neither is $\hat{\theta}^{\dagger}_d$ unbiased, even if both constituent estimators are unbiased under the appropriate conditions.

The relative strengths of the two estimators are exploited more effectively by combining them, as

$$\tilde{\theta}_d = (1 - b_d)\hat{\theta}_d + b_d\hat{\theta}\,, \tag{2}$$

especially when the districts can be regarded as exchangeable. The district specific coefficients b_d are set so as to minimize $\mathrm{MSE}(\tilde{\theta}_d;\, \theta_d)$, although a different objective (optimization) could be pursued instead.

The composite estimator $\tilde{\theta}_d$ exploits the similarity of the districts. When b_d is determined with precision, $\mathrm{MSE}(\tilde{\theta}_d;\, \theta_d)$ is smaller than both $\mathrm{MSE}(\hat{\theta}_d;\, \theta_d)$ and $\mathrm{MSE}(\hat{\theta};\, \theta_d)$ because the direct and pooled estimators are the extreme choices in Equation (2), corresponding to $b_d = 0$ and $b_d = 1$, respectively.

In most applications, $\hat{\theta}$ is a linear function of the data and $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2 + \cdots + c_D\hat{\theta}_D$. This motivates a more general form of Equation (2),

$$\tilde{\theta}_d = (1 - b_d)\hat{\theta}_d + \sum_{d' \neq d} b_{d',d}\hat{\theta}_{d'}. \tag{3}$$

When spatial correlation or some other form of dependence or *similarity* structure among the districts is present, it can be exploited similarly. In Equation (2), each district $d' \neq d$ is treated on an equal footing, being equally relevant (similar) to estimating θ_d. If the spatial neighbours of district d are more similar than the districts further afield, they can be assigned relatively greater weights in the composition [Equation (3)].

The optimal coefficients b_d and $b_{d',d}$ usually depend on some parameters, and so the coefficients are estimated. The analyst need not derive their estimators explicitly; the composition in Equation (2) or Equation (3), and its various generalizations that adjust for regressors, nonlinear link function, and the like, are direct byproducts of fitting models relevant to the problem of estimating $\tilde{\theta}_d$. These (linear) hierarchical models have separate representations for the within- and between-district variation:

$$f\{\mathrm{E}(y_{id}|\delta_d)\} = x_{id}\beta + z_{id}\delta_d; \tag{4}$$

where f is the link function and x and z are the sets of variables associated with regression and between-district variation, respectively. In the standard setting, z is a subset of the variables in x, z contains the intercept and with each interaction (product) variable the constituent variables ('main effects') are also included, in x and z, as applicable. These rules follow from the invariance of the model with respect to linear transformations.

The between-district variation is implied by the right hand side of Equation (4), and the conditional distribution of y_{id} may involve some parameters, such as the variance σ^2_{W} in models with the usual normality assumptions; $f(u) = u$ and

$$(y_{id}|\delta_d) \sim \mathcal{N}(x_{id}\beta + z_{id}\delta_d,\, \sigma^2_{\mathrm{W}}). \tag{5}$$

A model is formulated for δ_d. In the simplest specification, it is a random sample from a multivariate normal distribution with mean 0; the distribution applies to the districts. Associating districts with randomness is a logical inconsistency. In the frequentist viewpoint, randomness is a reference to the way hypothetical replications of the data-generating and estimation processes would be conducted. In the data generating process, the characteristics of any given district d, described by δ_d, would differ from one replication to next. For example, the rate of use of a particular medical service in a given district d would vary around the national rate across the replications. This is in conflict with the viewpoint of a typical survey analyst whose interest is in the variation over the replications of the sampling and estimation processes in a population that is 'frozen' or observed at a fixed time point.

A constructive way of resolving this contradiction is by agreeing that some invalid models are very useful for inference. Indeed, random effect models have transformed the study of spatial structures and discredited all forms of direct and pooled estimation, as well as their mixtures. In this context, we emphasize that a valid (or correct) model is a prerequisite for efficient estimation only *asymptotically*. For finite samples, it may be advantageous to incur a bias by using a submodel of the 'correct' model, and enjoy the reduced sampling variance *vis-à-vis* a 'correct' model.[1] Random effect models are more parsimonious than their fixed effect (ANCOVA) counterparts, because the D district level deviations are represented by a single parameter, a variance. When z in Equation (4) has r components, the district level deviations are described by D r-variate vectors in ANCOVA (Dr parameters), and an $r \times r$ variance matrix, involving $r(r+1)(1/2)$ parameters, in the corresponding random coefficient model. Usually $r \ll D$.

References to asymptotics may be appropriate for estimation of the *global* parameters in Equation (4), such as β and the residual variance σ^2_{W}, but not to local ones, such as the deviations δ_d. The between-area variance matrix $\Sigma_B = \mathrm{var}(\delta_d)$, or its univariate version, σ^2_{B}, are borderline cases, because the effective sample size for them is D or smaller. The small sample nature of the problem of estimating θ_d implies that estimation and transformation are not commutative. That is, if $\tilde{\theta}_d$ is an efficient estimator of θ_d, then $g(\tilde{\theta}_d)$ is not necessarily an efficient estimator of $g(\theta_d)$ for a nonlinear function g. Also, a nonlinear summary of the population quantities $\theta_d, S(\theta_{\mathcal{D}})$, is not estimated efficiently by the same summary of the efficient estimators of $\theta_d, S(\tilde{\theta}_{\mathcal{D}})$. Here, $\mathcal{D}$ is used as a collective index for the districts.

We assume that an appropriate model of the form [Equation (4)] has been identified and its parameters estimated efficiently. The coefficients b_d and $b_{d',d}$ in Equation (3) are functions of the model parameters, so they are *estimated* as $\hat{b}_d$ and $\hat{b}_{d',d}$, respectively. The coefficients are nonlinear functions of the model parameters. For example, with the two level model

$$y_{id} = x_{id}\beta + \delta_d + \varepsilon_{id}\,,$$

the mean deviation $\delta_d = \overline{y} - \overline{x}\beta$ would be estimated using $b_d = 1/(1 + n_d\rho)$, where n_d is the sample size of district d, and ρ is the variance ratio, $\rho = \mathrm{var}(\delta)/\mathrm{var}(\varepsilon)$. In this and many other settings, b_d and $b_{d',d}$ would yield efficient estimators of the target quantities θ_d *if* the model parameters were known. In practice, b_d and $b_{d',d}$ are estimated with nontrivial sampling variation, and the impact of this uncertainty on the efficiency of $\tilde{\theta}_\mathrm{d}(\hat{b}_d)$ is difficult to evaluate. An analyst using a software package, or its computational algorithm, as a black box has no opportunity to gain any appreciation of this problem. Two generic approaches to addressing this problem are to frame it in terms of missing information (as in the EM algorithm[2]), and to err on the side that has less severe consequences. For example, it is usually preferable to underestimate b_d because we increase the chances that the result is more efficient than the direct (unbiased) estimator $\hat{\theta}_d$. At the same time, however, the gains in precision may be reduced for most of the districts.

When the model parameters, a vector ξ, are regarded as missing data, a more efficient way of estimating the coefficients b_d and $b_{d',d}$ is by their conditional expectations given the data and the distribution of the parameter estimators. This has a straightforward interpretation as the E-step of an EM algorithm, in which the complete-data analysis evaluates $\tilde{\theta}_d$. Evaluation of the expectation in the E-step may involve multidimensional integrals; a complex task, despite the considerable progress made in recent years.[3] A computationally simpler approach is motivated by multiple imputation. Instead of the estimated values of the model parameters, $\hat{\xi}$, we generate *plausible* values of the parameters, $\tilde{\xi}$, as random draws from their estimated (joint) sampling distribution. In the equivalent Bayesian formulation, draws are made from the (joint) posterior distribution of the model parameters. The plausible parameter values generate plausible sets of coefficients $\tilde{b}_d$ and $\tilde{b}_{d',d}$, from which the estimator $\tilde{\theta}_d$ is evaluated as in Equation (3) or similar. This process is replicated several times, yielding estimates $\tilde{\theta}_d^{(m)}$, $m = 1, \ldots, M$. The multiple imputation estimator of $\tilde{\theta}$ is defined as their mean;

$$\tilde{\theta}_d^{\mathrm{MI}} = \frac{1}{M}\sum_{m=1}^{M} \tilde{\theta}_d^{(m)}\,.$$

Suppose an estimator of the sampling variance of $\tilde{\theta}_d$ is available, say, $\hat{s}_d^2$, that would be unbiased *if* the model parameters were known. Then the sampling variance of $\tilde{\theta}_d^{\mathrm{MI}}$ is estimated with small or no bias by

$$\tilde{s}_{d,\mathrm{MI}}^2 = \frac{1}{M}\sum_{m=1}^{M} \hat{s}_{d,m}^2 + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}\left(\tilde{\theta}_d^{\mathrm{MI}} - \theta_d\right)^2 ,$$

where $\hat{s}_{d,m}^2$ is the estimate of the sampling variance based on the *m*th set of plausible values of the model parameters. A technical condition attached to this statement is that the sampling variance of $\hat{s}_d^2$ is of smaller order of magnitude than s_d^4. For theoretical background, see Rubin.[4,5] The advantage of this approach is that little programming is required beyond implementing the algorithm that is efficient when the sampling variation of the model parameter estimators can be ignored. However, the standard output of some packages, quoting estimated standard errors, is not sufficient; the entire (estimated) sampling variance matrix is required, so that the correlation structure of the estimators is appropriately reflected.

In practice, we cannot establish that the model applied is appropriate. At best, we can identify any conflicts with the model assumptions. The model selection process is usually ignored, leading to gross biases.[6] A proposed solution is motivated by composite estimation.[1] Note that application of model checking and diagnostic procedures also leads to mixtures of *single model based* estimators if their result is a revision of the model or of the dataset. The uncertainty entailed in these procedures is generally ignored, assuming, in effect, that a replication of the data generating, modelling and model checking processes would result in the same conclusion. The consequent distortion permeates to the estimator of θ_d. An effective way of combatting this problem is by watering down the influence of the model in estimating θ_d. By reducing the coefficients b_d and $b_{d',d}$, we increase the reliance on the direct estimator $\hat{\theta}_d$; its properties are easier to establish, and are usually associated with more confidence. By relying more on $\hat{\theta}_d$ we lose the optimality *if* the model is appropriate, but protect our inferences from a distortion when it is not. The reduction of the coefficients b has to be set subjectively, reflecting our concerns about the inappropriateness of the model used and the process of selecting it. At the extreme, if the coefficients are reduced radically, the result differs only slightly from the direct estimator and the potential of exploiting similarity across the areas, variables, time and other factors is almost completely lost.

One unresolved matter stemming from the 'inappropriate' use of hierarchical models for 'fixed' areas is that the MSE of $\tilde{\theta}_d$ is estimated with bias; the quantity estimated without bias is the average MSE over the replications with *varying* δ_d. This can be interpreted as a form of averaging over areas with similar representation in the sample. Ideally, the sampling distribution should be estimated for the district on its own. For example, the deviation δ_d in Equation (5) is estimated by

$$\tilde{\delta}_d = \bar{e}_d \frac{n_d \rho}{1 + n_d \rho} ,$$

where $\bar{e}_d = \bar{y}_d - \bar{x}_d\hat{\beta}$ is the average of the residuals in area d. Ignoring the error in estimating β, the MSE of $\tilde{\delta}_d$ is

$$\begin{aligned}\mathrm{MSE}(\tilde{\theta}_d; \theta_d | \theta_d) &= \frac{\delta_d^2}{(1 + n_d\rho)^2} + \frac{\sigma_{\mathrm{W}}^2}{n_d}\frac{n_d^2\rho^2}{(1 + n_d\rho)^2} \\ &= \frac{\delta_d^2 + \sigma_{\mathrm{B}}^2 n_d\rho}{(1 + n_d\rho)^2};\end{aligned}$$

its average, obtained by replacing δ_d^2 with its area level expectation σ_{B}^2, is the more familiar expression $\sigma_{\mathrm{B}}^2/(1 + n_d\rho)$. The dependence on δ_d is very inconvenient because the MSE of the estimator (of δ_d) depends on the target itself. A naive estimator of δ_d^2 is obtained by replacing δ_d^2 by $\tilde{\delta}_d^2$, although δ_d^2 could be estimated by a composition directly.

3 Inferences with selected targets

The models considered thus far deal with a ‘frozen’ population and *a priori* set targets of estimation. In the context of disease mapping for a given domain with a fixed partition into districts, this corresponds to a particular time point. The distribution of a studied phenomenon changes over time, so the results of the analysis for a particular time point give us only a snapshot. Extrapolation from the time point to the future (or past) is appropriate only to the extent of the temporal correlation (inertia, or similarity). In brief, variation is present not only across the districts but also in time. This implies that the choice of the time point for the observations (data collection) and analysis is not innocuous. When the time point is selected uninformatively, such as by a date specified in advance, we obtain an ‘unbiased’ snapshot of the domain. When the date is selected after an event or a decision made in response to some observations about the phenomenon relevant or related to the data, the time point is selected informatively, and the snapshot is biased. This is not a problem when the time point is selected deliberately, and the selection is incorporated in the interpretation of all the results. However, secondary analysts may not be informed about the nature of the selection; then all their inferences are distorted. Quoting the time point is not helpful; the purpose that resulted in selecting the particular time point should be quoted. The purpose can be interpreted as the result of a process of selection of the target.

When the districts are enumerated, as when data are obtained from administrative registers, no selection issues arise. When the elementary data are collected by a survey, the sampling design is important; not the design that was planned, but the one that was realized. In particular, the realized design incorporates the process of nonresponse as well as imperfect coverage of the domain. The principles of sampling design are applicable not only in settings with explicitly defined populations (domains), but also whenever inferences are intended for a collection of settings, such as time points, outcome variables or geographical units. In settings of monitoring or auditing, the process of observation and reporting should be regarded as a kind of highly informative sampling process, because a given dataset, together with its context, is submitted for analysis only after a careful inspection and deliberation.

Many problems in disease mapping, and small area estimation in general, relate to collections of quantities (one or a few for each district). This entails complexities of two kinds. First, efficient estimation of a quantity does not lead to optimal estimation of its nonlinear transformation, as pointed out earlier. More generally, the pattern observed among the estimates $\tilde{\theta}_d$ may be systematically (over hypothetical replications) different from the pattern among the population quantities θ_d. As a simple example of this problem, suppose each of a large number of districts is represented in a survey by only two observations. The within-district variance is substantial, so the direct estimator of each θ_d is very inefficient. Suppose the between-district variation is moderate. An empirical Bayes or a related method estimates each θ_d by a combination of the direct estimator and a statistic based on the rest of the survey data. The latter is assigned much greater weight because it is preferable to incur the bias induced by the auxiliary (external) information than the substantial sampling variation of the direct estimator. As a consequence, the estimates $\tilde{\theta}_d, d = 1, \ldots, D$, are dispersed much less than the underlying quantities θ_d. In contrast, the direct estimates $\hat{\theta}_d$ are dispersed much more than the quantities θ_d, because their variation is composed of the two (independent) components: estimation errors $\hat{\theta}_d - \theta_d$, and differences among the population quantities θ_d. A solution of this problem is often sought by *smoothing*. The composition in Equation (3) can be regarded as an example of smoothing, or even as its definition, especially if the coefficients $b_{d',d}$ are a function of the distance between districts d' and d. The appropriate extent of smoothing is difficult to set, unless we have a preconceived idea of how smooth the set of estimates $\tilde{\theta}_d$, or the map based on them, should be.

Suppose an appropriate extent of smoothing has been applied. How well supported is the inference that a feature identified in the smooth map (among the smoothed estimates) is also a feature that would be observed among the population quantities $\theta_{\mathcal{D}}$? Outliers, a pattern, breaks in the pattern, smoothness or, in general, a particular summary $\mathcal{F}(\hat{\theta}_{\mathcal{D}})$, can all be regarded as features. A single map cannot inform us whether a particular feature is due to the sampling variation or reflects the presence of the feature in the population. That is, whether $\mathcal{F}(\hat{\theta}_{\mathcal{D}}) \doteq \mathcal{F}(\theta_{\mathcal{D}})$ or, more precisely, acknowledging that $\mathcal{F}(\hat{\theta}_{\mathcal{D}})$ is a random object, whether $\mathrm{E}\{\|\mathcal{F}(\hat{\theta}_{\mathcal{D}}) - \mathcal{F}(\theta_{\mathcal{D}})\|^2\} < \Delta$, for a suitable metric $\| \cdot \|$ and threshold $\Delta > 0$.

The uncertainty about a feature of the map cannot be represented by a simple object, such as a number (akin to the conventional standard error). A comprehensive way of representing such uncertainty is by drawing a few maps of plausible sets of values of θ_d, $d = 1, \ldots, D$. The mechanics of generating plausible values has been described earlier. Care has to be taken in their generation to reflect not only the sampling variances of the quantities involved, but also their sampling correlations. For example, when the estimators of θ_d for pairs of neighbours are highly correlated, their values will be almost linearly related across the sets of plausible values. With plausible maps drawn in this way, a feature can be reported as a characteristic of the population when it is present in (almost) all plausible maps. (These are difficult to present in a publication because they take up several pages.) The plausible maps can be instructive for learning about the kinds of features that can occur by chance; in other words, about our propensity to observe patterns where there are none, and also about patterns that are missed when only one map is inspected. Only the replicable is remarkable.[7]

4 Studying extremes

For many phenomena related to health care and epidemiology, the universe studied is the residents of a domain (country) over a period of time. Often the relevant data, an enumeration or its compact summary, are available but the need for a specific analysis arises from an unusual or unexpected observation of the domain. In such a setting, the selection of the time point is highly informative, and the inferences based on the subsequent analysis cannot be interpreted as a description of the domain at a random snapshot. Further, the analysis may be triggered by an observation related to one or a few districts. Then the inferences about the district, based on an analysis that treats these districts on a par with the others, are distorted. They are not inferences about district d, but about the extreme district in the domain, or the district selected by a particular process. The conditioning on the 'trigger' for the analysis is essential.Without it, we are asking whether a horse that appears to have won the race has really won it! With the race in progress, the betting odds rapidly lose their relevance.

This analogy carries over to many other settings. For example, a court of law may hear evidence that, given honesty of the accused, the chances of an outcome at least as extreme as the one documented about the accused is one in a thousand. Such a statement and its connotation are appropriate in the alien setting when a person randomly drawn from the relevant population is accused. Then the conditional probability of an inappropriate conviction, $P(\text{conviction}|\text{innocence})$, based solely on this probabilistic statement, is indeed 0.001. Note that the probability of a correct decision, $P(\text{conviction}|\text{guilt})P(\text{guilt}) + P(\text{acquittal}|\text{innocence})P(\text{innocence})$, is not 0.999. However, the process of selection of the accused is highly informative. If the accused is carefully selected by a criterion or an informal process that is closely related to the statistic or summary used in deriving the figure of 1 : 1000, then the court is presented a near-tautology, and the jury is asked to convict the defendant, in effect, because the defendant has been accused. In a population of thousands of honest and law abiding people, there are a few 'unlucky' members who display an *a priori* specified pattern that has probability 0.001 under the null hypothesis of 'universal honesty'. If some of these 'unlucky' members (and no others) are accused, based on the '1 : 1000' or a related statement of fact, which is then repeated in the court, *every* conviction is unjust.

An event in the population has probability 0.001 but, in a particular class of court trials, the same event has a much higher probability, because of the careful selection of cases presented to it. Just like probabilities (Bernoulli distributions), properties of distributions are also highly contingent on conditioning.[8] The distribution of an estimator in one setting (selected *a priori*) may be very different from its distribution in another setting (when selected as a result of a data inspection). In other words, the process of target selection is informative and, for hypothesis testing, the null distribution of the test statistic should be adjusted accordingly. For example, if the selection points to the area with the lowest of the values of $\hat{\theta}_d$, $d = 1, \ldots, D$, the appropriate hypothesis is that the *minimum* of the values of θ_d displays a particular feature. Testing this hypothesis might provide evidence that the minimum of θ_d is exceptional. Strictly speaking, the area d that is then identified naively, as the area with the smallest $\hat{\theta}_d$, need not be the exceptional area, because the identification is subject to uncertainty, additional to that entailed in the hypothesis test. A recent prominent example of an *a posteriori* selected target is the analysis of the Bristol inquiry into the death rates in heart surgeries of children.[9] The area with the minimum θ_d may still not be an appropriate target that would incorporate the target selection process, because the identified feature is not the minimum, but an *outstanding* minimum in a two way array of estimates $\hat{\theta}_{d,t}$, where t denotes occasions. Simply, the more the occasions the greater the chance that the minimum of the estimates stands out for *some* occasion. The appropriate hypothesis in such circumstances is about neither *the* area nor *the* time point, but about *an* area at *a* time point, and the hypothesis is tested only when the alarm is raised by a scheduled audit or chance inspection of the outcomes.

We use the term *personalization* for (inappropriate) labelling of a specific district, time point, or the like, as the target of inference. If we personalize a target, the process of its selection has to be accounted for in the analysis. Without personalization, the relevant inferences are about extremes, although these also require a careful specification.

A typical trigger for conducting an analysis is set off by a well founded suspicion of an active external agent that is the cause of a substantial change in the value of θ_d for one or a group of districts. The remit of a subsequent analysis is to confirm this by assessing the evidence that the district(s) stand out among the rest. Although a lot of progress in modelling extremes has been made in the recent years,[10,11] problems involving them remain poorly understood and available methods are not applicable universally. With abundant computing power, simulations can come to the rescue. We specify the extent of outlying as a feature, and pose the question:

> How likely is such a feature when there are no external agents?

We simulate data sets for the country and its districts from a model that assumes no external agents, and apply the same criterion for identifying the same feature as was applied with the 'real' data. Absence of the feature in most simulated data sets is evidence that the feature is a result of an external agent (departure from the model), because without it the feature is unlikely to occur. A closely related issue arises in the context of outliers and model diagnostics.[12,13] The feature itself should not be personalized; all features that might trigger an investigation should be considered.

A difficulty in this simulation approach is the specification of the model from which the artificial data sets are to be generated. Ideally, it would be *the* model that describes the studied phenomenon in the domain in the recent past, so that not much extrapolation takes place and sufficient hindsight is available to confirm the absence of any external agent in that period. Of course, the model cannot be identified with all its details, so a fitted or plausible model has to be used instead. The robustness of the conclusions is established by a sensitivity analysis, carrying out the simulations for several plausible models.

The population quantities $\theta_{d,t}$ differ both in time and across areas. If we focus on a single time point t, $\theta_{d,t}$ for an area should be regarded as remarkable not when it differs from the national quantity θ (significantly, or by a wide margin), but when it stands out among the values of $\theta_{d',t}$ for the other areas d'. An obvious target selection process identifies the area $d^\dagger$ with the highest estimate $\tilde{\theta}_{d^\dagger}$. Acknowledging this process, the appropriate null hypothesis relates to the *maximum* of population quantities θ_d. A common mistake is that the area $d^\dagger$ becomes personalized, and hypotheses are formulated about the area $d^\dagger$ as if it were identified *a priori*. In practice, an area is identified not solely based on $\hat{\theta}_d$, but on several informal comparisons, and so the target selection process defies a formal description. The inability to specify the conditioning involved does not justify the reference to the unconditional distribution of the estimator $\hat{\theta}_{d^\dagger}$.

As extremes are identified by comparison with the 'usual', it is essential to study the 'usual'. This entails estimating the extent of between-district variation when the absence of an external agent is confirmed, as well as within-district variation over time. This would enable a clearer formulation of what should be regarded as exceptional and not a result of the myriad of background processes that generate the usual state differences among the districts. Simulations from the usual state can inform us about the extent of outlying and other outstanding features in the data that can reasonably be expected.

5 Intervention under uncertainty

Establishing the presence of an external agent, in a particular district or region, is a signal for implementing special measures to locate and eradicate it, or to reduce its influence to a minimum. Such measures are often extensive and involve substantial expenditure of labour, time, finance and community goodwill. The associated risk includes the loss of political and professional reputation and confidence in the authority implementing the measures, especially when it relies on the co-operation of the public. The application has an expected impact, say, the reduction of the value of θ_d back to the 'national' or regional background level, so that it would no longer be exceptional. The two kinds of incorrect decisions, combatting the agent when it is not present and assuming its absence when it is present, lead to huge costs that are usually of unequal magnitude. The costs need not be only in terms of monetary funds, but more generally in values agreed by the parties involved, including representatives of the public at large and the authorities in the districts concerned. They may include harm, in the short or long term, to the health and well being of the residents, its threat leading to inconvenience, hindrance to economic and leisure activities and transportation, loss of business income and the cost of preventative measures. Uncertainty about the impact of the measures is another component of the costs that has to be considered. No evidence of any impact is an insufficient criterion. Evidence of no greater than negligible impact is desirable.

With incomplete information in the statistical analysis, the presence of an external agent is not established with certainty, and so the two kinds of errors *may* be committed. Limiting the conditional probability of one kind is the traditional approach related to hypothesis testing. We regard it as outdated and irrelevant because it ignores the relative penalties associated with the two kinds of erroneous decisions. Similar criticism of hypothesis testing has been voiced with a different perspective and applied to a different setting,[14] but the arguments presented readily carry over to our setting.

We illustrate the point on two scenarios. First, suppose the cost of intervention, the result of incorrectly assuming that an external agent is present, is much smaller than the damage caused by the agent when it is present. In this setting, we should play safe and intervene at the slightest suggestion of its presence, because intervention costs next to nothing. When the agent causes little damage, but the intervention is costly, we should think twice, because we derive little benefit following a lot of expenditure. Without being informed about the relative costs, the strength of evidence about the presence of the agent is not an appropriate index for decision making. A hypothesis test or a similar criterion would suggest taking a particular action without being informed about its potential costs and benefits. A more intelligent and coherent approach incorporates the costs in the statistical analysis.

We consider the simplified scenario in which, at any time point, there are three courses of action: 1) to do nothing, 2) to implement an intervention or 3) to seek additional information. These activities cost $c_1 = 0$, $c_2 > 0$ and $c_3 < c_2$ units, respectively. The penalty for doing nothing when an intervention is appropriate is C_1, and for intervention when it is not appropriate it is C_2. When additional information is sought, there is a penalty C_{31} when this is unnecessary and a much greater penalty C_{32} when intervention should be applied. In practice, these costs are better formulated as per unit of time (say, per day), although the costs need not be linear functions of the time elapsed. Further, the costs and penalties may not only depend on the extent of the presence of the agent, but also on the value of the key parameter, θ. This caters for the situation when an intervention is applied, and an investigation after the crisis establishes whether there was any good cause for the intervention and, if there was none, how obvious it was, should have been or may have been.

We assume that the following decision rule is adopted: if $\hat{\theta} < \theta^{(12)}$, no action is taken, if $\hat{\theta} > \theta^{(23)}$, intervention is applied, and otherwise, if $\theta^{(12)} < \hat{\theta} < \theta^{(23)}$, further information is sought. Suppose the value of the parameter of interest θ is certainly in the range (θ_L, θ_U). A Bayesian may use a prior distribution for θ instead. For a value $\theta_\dagger \in (\theta_L, \theta_U)$, we evaluate the distribution of $\hat{\theta}$ assuming $\theta = \theta_\dagger$, and the expected loss due to the activities and associated penalties, when applicable. The outcome of this exercise is a function of θ, with additional parameters $\theta^{(12)}$, $\theta^{(23)}$ and parameters that represent the choice of the estimator $\hat{\theta}$. The choice of the thresholds $\theta^{(12)}$ and $\theta^{(23)}$ is based on these penalty curves. Either each curve is summarized, for example, by its mean value, or a more complicated functional is used that better reflects the preferences of the parties involved.

At first sight, this appears to be a computationally extensive procedure. It is extensive by the standards of the computing technology of a few decades ago; but nowadays, it does not represent an undue computational burden. More importantly, it is easy to program, because usually its most complex elements are repeated evaluations of the estimator $\hat{\theta}$ and of the penalty curve. In principle, the costs of computing, including the relevant expertise and manpower, could be included in this model. However, these costs are much less in comparison with what is at stake when the intervention considered is extensive and expensive.

A likely reason for why statistics is not involved integrally in such decisions is that uncertainty, that is, not being in possession of complete information, or not having the analytical ability to process it, is looked upon as a weakness. It certainly is, but its denial is more harmful than its admission because it compounds the problem and does not encourage efforts to reduce it by searching for or inviting the provision of relevant information. The optimal strategy is not to try and eradicate the uncertainty, but to reduce it, and carefully weigh the benefits of such a reduction – a more appropriate decision – against the costs (including time) of collecting the relevant information.

6 Conclusion

We are accustomed to reporting estimates and the associated estimated standard errors, and quoting their *unconditional* properties, such as (small or) no bias and (asymptotic) efficiency when applying a maximum likelihood estimator. In some settings, selective reporting is criticized as poor practice, implying that we should avoid it. Disease mapping is an example of selective reporting in which the selection is essential and often appropriate. Although not proposing any universal solutions, we have highlighted the need to consider the process of target selection because it has a profound impact on the properties of the estimators, in parallel with the impact of conditioning on the value of a probability.

In many ongoing data collection exercises, involving surveillance or monitoring of social, epidemiological and geophysical phenomena, statistical expertise is often involved only after an unusual feature has been identified. Whether such a feature is genuinely unusual can be assessed with rigour only by incorporating the information about the process of identifying the feature in the first place, because such a process is highly informative in that it has an impact on the distribution of the estimator consequently applied.

Similarly, the decision about the measure taken after the analysis, no action or intervention, should be integrated in the analysis itself, because the decision has to be informed by the relative sizes of the penalties (harm done) when an inappropriate action is taken. Conventional analyses (aim to) minimize the probabilities of making inappropriate decisions. This is an extremely ineffective strategy when the penalties are of different orders of magnitude.

Most of the standard statistical theory deals with estimators, and with inference in general, without any conditioning. The processes that we should condition on, such as target selection, are often rather complex, defying any straightforward description. There are no universally applicable theoretical methods for incorporating them in the analysis. Simulation of scenarios offers us a low tech, even if computationally extensive, way of studying the impact of the process of target selection, and widening our horizons from operating with unconditional distributions of estimators to studying their conditional distributions given the processes of identifying them as being relevant.

Acknowledgements

The article was prepared and the research for it conducted while the author was an Academic Visitor at the Institute of Information Science and Technology, Massey University, Palmerston North, New Zealand. Hospitality of the Institute and its stimulating academic environment are acknowledged.

References

1 Longford NT. An alternative to model selection in ordinary regression. *Statistics and Computing* 2003; **13**: 67–80.

2 Dempster AP, Laird NM, Rubin DB. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 1977; **39**: 1–38.

3 Lesaffre E, Spiessens B. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics* 2001; **50**: 325–35.

4 Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley and Sons, 1987.

5 Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**: 473–89.

6 Draper DN. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B* 1995; **57**: 45–98.

7 Gelman A, Price PN. All maps of parameter estimates are misleading. *Statistics in Medicine* 1999; **18**: 3221–34.

8 de Finetti B. *Theory of probability. A critical introductory treatment*. Volume 1. Translated from Italian by Machí, A, Smith A. Chichester: John Wiley and Sons, 1974.

9 Spiegelhalter DJ, Aylin P, Best NG, Evans SJW, Murray GD. Commissioned analysis of surgical performance using routine data: lessons from the Bristol inquiry. *Journal of the Royal Statistical Society: Series A* 2002; **165**: 191–231.

10 Coles SG, Dixon MJ. Likelihood-based inference for extreme value models. *Extremes* 1999; **2**: 5–23.

11 Ledford AW, Tawn JA. Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B* 1997; **59**: 475–99.

12 Gelman A, Carlin JB, Stern H, Rubin DB. *Bayesian data analysis*. London: Chapman and Hall, 1995.

13 Longford NT. Simulation-based diagnostics in random coefficient models. *Journal of the Royal Statistical Society: Series A* 2001; **164**: 259–73.

14 Lindley DV. Decision analysis and bioequivalence trials. *Statistical Science* 1998; **13**: 136–41.