NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

01-21

A Few Statistical Principles for Data Science

Noel Cressie

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

Aust. N. Z. J. Stat. 2020 doi: 10.1111/j.1467-842X.XXX

A few statistical principles for data science

Noel Cressie[1]*

University of Wollongong

Summary

In any other circumstance, it might make sense to define the extent of the terrain (Data Science) first, and then locate and describe the landmarks (Principles). But this data revolution we are experiencing defies a cadastral survey. Areas are continually being annexed into Data Science. For example, biometrics was traditionally statistics for agriculture in all its forms but now, in Data Science, it means the study of characteristics that can be used to identify an individual. Examples of non-intrusive measurements include height, weight, fingerprints, retina scan, voice, photograph/video (facial landmarks and facial expressions), and gait. A multivariate analysis of such data would be a complex project for a statistician, but a software engineer might appear to have no trouble with it at all. In any applied-statistics project, the statistician worries about uncertainty and quantifies it by modelling data as realisations generated from a probability space. Another approach to uncertainty quantification is to find similar data sets, and then use the variability of results between these data sets to capture the uncertainty. Both approaches allow 'error bars' to be put on estimates obtained from the original data set, although the interpretations are different. A third approach, that concentrates on giving a single answer and gives up on uncertainty quantification, could be considered as Data Engineering, although it has staked a claim in the Data Science terrain. This article presents a few (actually nine) statistical principles for data scientists that have helped me, and continue to help me, when I work on complex interdisciplinary projects.

Key words: Applied statistics; hierarchical statistical models; measurement error; regression; spatial statistics

1. Does your analysis have 'error bars'?

Science and Engineering have long been held separate by our universities and learned academies. Broadly speaking, Science is concerned with 'why,' and Engineering is concerned

*Author to whom correspondence should be addressed.
[1]National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong
Wollongong NSW 2522, Australia
Email: ncressie@uow.edu.au

Acknowledgment. This research was supported by the Australian Research Council under Discovery Project DP190100180. The Sydney Business School, University of Wollongong, generously provided an environment where most of this article was written. The referees have generously given comments that have helped me clarify the exposition of my nine principles. My thanks go to Dr B. Maloney for his skillful technical assistance. I would like to express my appreciation to Dr J. Wong for insightful discussions on this and other material.

Prepared using anzsauth.cls *[Version: 2018/01/30 Version 9]*

with 'how.' Data Science seems to include both (Cressie 2020), and sometimes it may be hard to find the science in a Data Science solution. For example, an idea is born on how a particular data set could be analysed, perhaps with only a vague idea of the question being answered. Then algorithms and software are developed and, in some cases, the analysis results in superior performance as judged by figures-of-merit applied to that data set, which leads to a publication. Are the results extendible and do they have 'error bars'?

Usually figures-of-merit are concerned with the *efficiency* of a method. However, *validity* should be established first, and error bars allow this to be done. No matter how efficient a prediction method is, if its error bars define a nominal 90% prediction interval but the true quantity (i.e., the predictand) is only contained in these intervals 60% of the time, the method lacks validity. This important figure-of-merit is called *coverage* and is analogous to one in the medical literature called specificity.

This article is aimed at data scientists who want their error bars to be based on probabilities and their coverages to be (approximately) as stated. There are data scientists who have error bars that are based on empirical distributions derived from ensembles of like data sets or from partitioning a single data set into like subsets. While the principles I present are for data scientists who quantify uncertainties using probabilities, they could act as a guide for both types. If your analyses include uncertainty quantifications, then you may find some of the principles in the following sections useful for hypothesis generation, for climbing the knowledge pyramid from data to information to knowledge, and for scientific inference.

2. My principal principles

When I was planning this article, I hovered between the words 'laws,' 'rules,' 'principles,' and 'guidelines' to establish some landmarks in Data Science. Laws should not be broken, exceptions prove the rule, and guidelines prevent accidents. But principles should make the analysis go better without interdicting other approaches: Improvising on something Groucho Marx once said, if you don't like my principles, you may have others you like better . . .

A data scientist might see variability between strata and within strata. A statistical scientist uses probability distributions to capture some or all of this variability. For example, probability densities would typically be used to model within-strata variabilities, and these densities may be different across strata in order to capture the between-strata variability. In many, many cases, those strata differences are modelled via a regression model, leaving the errors around the regression line (i.e., the within-strata variability) to be described by a probability distribution, often a Gaussian (or sometimes called normal) distribution.

Prepared using anzsauth.cls

In my opinion, the worst word in the Statistics lexicon, which happens to have four letters, is 'mean.' It is the first moment of a probability distribution, and it is also used to describe the average of a collection of numbers, leading to many misunderstandings in Data Science. To data scientists who capture variability through the empirical distributions of summary statistics from similar data sets, the average is their first moment, so *they* could legitimately call it a mean of their empirical distribution. To avoid confusion, it will be henceforth assumed in this article that variability is captured by *probability* distributions, and a mean *iss* (is and only is) the first moment of that distribution.

In this section, five general statistical principles are presented. Different areas of study will have their own and, in Section 3, I present three principles that are important in Spatial Statistics. In an epilogical section (Section 4), I present one more, the principle of respecting units, which makes nine principles in total.

2.1. All models are wrong . . . but some are wrong-er than others

In my work on remote sensing of carbon dioxide, data $\boldsymbol{z}$ from a single sounding is a more-than-2,000-dimensional vector of radiances observed at wavelengths in the near infra-red part of the spectrum. Cressie (2018) contains a literature review that features the relationship between $\boldsymbol{z}$ and the state of the atmosphere $\boldsymbol{y}$ (a 40-to-50 dimensional vector) via the statistical model,

$$\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{y}) + \boldsymbol{\varepsilon}\,, \tag{1}$$

where unobserved $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are statistically independent; $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon)$, where $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ refers to a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $\boldsymbol{f}$ is a known vector of nonlinear forward models obtained from atmospheric physics/chemistry; $\boldsymbol{\Sigma}_\varepsilon$ is a covariance matrix known from experiments on the remote sensing instrument, performed in the laboratory before the satellite is launched; $\boldsymbol{y} = \boldsymbol{\mu}_\alpha + \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\alpha)$; and $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\Sigma}_\alpha$ are the first two moments of the (hidden) atmospheric state $\boldsymbol{y}$. The goal is to predict $\boldsymbol{y}$ from data $\boldsymbol{z}$. In my experience, it is the specification of $\boldsymbol{\Sigma}_\alpha$ (the covariance matrix of the state of the atmosphere) where the statistical model can be seriously wrong (Cressie 2018).

Principle 2.1: *Establish a true model (TM), perhaps different from the scientist's working model (WM). Critically, compute the TM-distributional properties of the WM estimators.*

At the very least, this principle gives the data scientist an idea of the sensitivity of the model to misspecification. In the remote sensing example, the WM-based statistical analysis provides a predicted atmospheric state, $\hat{\boldsymbol{y}}_{WM}$, called a *retrieval.* However, the atmosphere does not care what WM was chosen; it acts under the true model (TM), and the principal

figures-of-merit that are calculated are (using obvious notation):

$$\begin{aligned} &\text{True bias: } \mathrm{E}_{TM}(\hat{\boldsymbol{y}}_{WM} - \boldsymbol{y})\,, \\ &\text{True uncertainty: } \mathrm{var}_{TM}(\hat{\boldsymbol{y}}_{WM} - \boldsymbol{y})\,. \end{aligned} \tag{2}$$

While $\hat{\boldsymbol{y}}_{WM}$ may be optimal under WM, it is not under TM, and the properties given by (2) will expose the potential weaknesses of a WM chosen for its simplicity or its computational convenience.

It is tempting (but wrong) for scientists to calculate instead:

$$\begin{aligned} &\mathrm{E}_{WM}(\hat{\boldsymbol{y}}_{WM} - \boldsymbol{y})\,, \\ &\mathrm{var}_{WM}(\hat{\boldsymbol{y}}_{WM} - \boldsymbol{y})\,, \end{aligned} \tag{3}$$

which are often facile calculations compared to (2). This was the case for NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite. Over time, empirical distributions of retrieval-error ensembles showed that bias and variance should have been calculated from (2), not (3). This is discussed by Cressie (2018), and an illustration of applying Principal 2.1 is given in Nguyen, Cressie & Hobbs (2019).

In the rest of this article, I shall assume that the WM and TM are one and the same. However, this principle is potentially applicable in all the examples below.

2.2. What you see is not what you want to get

There is an equivalent way to write the model (1) that shows how statistical scientists can naturally build complexity into their models:

$$\begin{aligned} \boldsymbol{z}|\boldsymbol{y} &\sim \mathrm{N}(\boldsymbol{f}(\boldsymbol{y}), \boldsymbol{\Sigma}_\varepsilon) \\ \boldsymbol{y} &\sim \mathrm{N}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)\,. \end{aligned} \tag{4}$$

This is a *hierarchical statistical model (HM)*, which is defined by layers of conditional probabilities (here two layers).

The top layer is the *data model* and can be written in abbreviated notation as $[\boldsymbol{z}|\boldsymbol{y}]$. It models what we see – the data (conditional on the process $\boldsymbol{y}$). The next layer is the *process model* that can be written as $[\boldsymbol{y}]$. It models what we want to get – the latent process, or state, hidden behind the data. In fact, in the remote sensing problem described in Section 2.1, the distribution of $\boldsymbol{y}$ is conditional on $\boldsymbol{w}$, the meteorology of the atmosphere. If need be, the process model could be modified to now consist of $[\boldsymbol{y}|\boldsymbol{w}]$ and $[\boldsymbol{w}]$. Then the HM would be made up of three levels, and the sequence of conditional probabilities would be $[\boldsymbol{z}|\boldsymbol{y}]$, $[\boldsymbol{y}|\boldsymbol{w}]$,

and $[\boldsymbol{w}]$. In the specific case of OCO-2 retrievals, the meteorological process, $\boldsymbol{w}$, is obtained from a numerical weather forecasting model and is considered known.

Principle 2.2: *Build statistical models conditionally, through a data model and a process model. Infer the unknown process from the predictive distribution.*

The predictive distribution is,

$$[\boldsymbol{y}|\boldsymbol{z}] = \frac{[\boldsymbol{z}|\boldsymbol{y}] \cdot [\boldsymbol{y}]}{[\boldsymbol{z}]}\,, \tag{5}$$

where for convenience any parameters $\boldsymbol{\theta}$ in $[\boldsymbol{z}|\boldsymbol{y}]$ and $[\boldsymbol{y}]$ are dropped from the notation (but they are still there). The HM components are featured in the numerator, which is equal to the joint distribution, $[\boldsymbol{z}, \boldsymbol{y}]$; and the denominator, $[\boldsymbol{z}]$, is simply a 'normaliser' to ensure that $[\boldsymbol{y}|\boldsymbol{z}]$ is a density (or probability mass) function that integrates (or sums) to 1. Equation (5) is Bayes' Theorem but, in this case, the unknowns are the state elements in $\boldsymbol{y}$, not the parameters $\boldsymbol{\theta}$. No prior distributions on $\boldsymbol{\theta}$ have been assumed in the making of this HM! However, a prior could be assumed or, pragmatically, an estimate $\hat{\boldsymbol{\theta}}$ could be obtained from $\boldsymbol{z}$.

Derivation of the predictive distribution in (5) can be problematic in many cases of practical interest, such as when $[\boldsymbol{y}]$ is a highly multivariate probability distribution of a complex scientific process. Computational methods, particularly Markov chain Monte Carlo (MCMC), are in constant development to allow realisations from $[\boldsymbol{y}|\boldsymbol{z}]$ to be simulated, from which distributional properties such as the predictive mean and the predictive quantiles can be used to predict the state $\boldsymbol{y}$ and the error bars, respectively.

The simplest special case of (4) is when the data (here, z) and the process (here, y) are univariate, and $f(\cdot)$ is the identity function:

$$\begin{aligned} z|y &\sim \mathrm{N}(y, \sigma^2_\varepsilon)\,, \\ y &\sim \mathrm{N}(\mu_\alpha, \sigma^2_\alpha)\,. \end{aligned}$$

Then $[y|z]$ is Gaussian, characterised by its first two moments:

$$\begin{aligned} \mathrm{E}(y|z) &= \{(1/\sigma^2_\alpha)\mu_\alpha + (1/\sigma^2_\varepsilon)z\}\{1/\sigma^2_\alpha + 1/\sigma^2_\varepsilon\}^{-1}\,, \\ \mathrm{var}(y|z) &= \{1/\sigma^2_\alpha + 1/\sigma^2_\varepsilon\}^{-1}\,. \end{aligned}$$

The predictor, $\hat{y} = \mathrm{E}(y|z)$, is a weighted combination of the data and the mean of the unknown state y, where the weights depend on the signal-to-noise ratio, $\sigma^2_\alpha/\sigma^2_\varepsilon$. These results can be generalised to the multivariate case with linear retrieval equations in (4), namely $\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{c} + \boldsymbol{K}\boldsymbol{y}$. In the case of OCO-2, $\boldsymbol{K}$ has more rows than columns, and the first two

moments of the predictive distribution, which is Gaussian, are:

$$\begin{aligned} \mathrm{E}(\boldsymbol{y}|\boldsymbol{z}) &= \boldsymbol{y}_\alpha + \boldsymbol{G}(\boldsymbol{z} - \boldsymbol{c} - \boldsymbol{K}\boldsymbol{y}_\alpha)\,, \\ \mathrm{var}(\boldsymbol{y}|\boldsymbol{z}) &= \{\boldsymbol{\Sigma}_\alpha^{-1} + \boldsymbol{K}^\top \boldsymbol{\Sigma}_\varepsilon^{-1} \boldsymbol{K}\}^{-1}\,, \end{aligned} \tag{6}$$

where $\boldsymbol{G} = \{\boldsymbol{\Sigma}_\alpha^{-1} + \boldsymbol{K}^\top \boldsymbol{\Sigma}_\varepsilon^{-1} \boldsymbol{K}\}^{-1} \boldsymbol{K}^\top \boldsymbol{\Sigma}_\varepsilon^{-1}$ is sometimes called the gain matrix.

From (6), it can be seen that the precision matrix (i.e., the inverse of the variance matrix), written $\mathrm{prec}(\cdot)$, of each component of the predictive distribution satisfies,

$$\mathrm{prec}(\boldsymbol{y}|\boldsymbol{z}) = \mathrm{prec}(\boldsymbol{y}) + \boldsymbol{K}^\top \mathrm{prec}(\boldsymbol{z}|\boldsymbol{y}) \boldsymbol{K}\,,$$

a result that holds when the matrix $\boldsymbol{K}$ is any size. This decomposition of precision demonstrates that, when going from $[\boldsymbol{y}]$ to $[\boldsymbol{y}|\boldsymbol{z}]$, the precision increases (i.e., the variance decreases). Moreover, the predictive mean is unbiased; that is, $\mathrm{E}(\mathrm{E}(\boldsymbol{y}|\boldsymbol{z})) = \mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu}_\alpha$.

In geostatistics, the importance of (5) is deeply misunderstood, because Matheron (1963) originally formulated kriging in terms of what he called a regionalised variable, $\{z(\boldsymbol{s}) : \boldsymbol{s} \in D \subset \mathbb{R}^d\}$, where $\mathbb{R}^d$ is d-dimensional Euclidean space and D is a subset of $\mathbb{R}^d$ with volume $|D| > 0$. In this formulation, the data model and the process model are collapsed into the single probability distribution, $[\{z(\boldsymbol{s}) : \boldsymbol{s} \in D\}]$. Then the goal is to predict $z(\boldsymbol{s}_0)$ from data $\boldsymbol{z} = (z(\boldsymbol{s}_1), \ldots, \boldsymbol{z}(\boldsymbol{s}_n))^\top$, for which the generic kriging predictor is $\mathrm{E}(z(\boldsymbol{s}_0)|\boldsymbol{z})$, the mean of $[z(\boldsymbol{s}_0)|\boldsymbol{z}]$.

However, when measurement of the process is taken into account, there is a HM that differentiates between the observations $\boldsymbol{z}$ and the underlying latent process $\{y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ and that is observed imperfectly through $\boldsymbol{z}$. The spatial trend and the spatial covariance function are defined in the process model, although they are estimated from the noisy data $\boldsymbol{z}$. Once the measurement error is included in the model, it is clear that geostatistics should do kriging using $\mathrm{E}(y(\boldsymbol{s}_0)|\boldsymbol{z})$ and not $\mathrm{E}(z(\boldsymbol{s}_0)|\boldsymbol{z})$; see Cressie & Wikle (2011, Section 4.1).

Consequently, much of the earlier geostatistics software did not make inference on $y(\boldsymbol{s}_0)$ but chose to make inference on $z(\boldsymbol{s}_0)$ where the measurement error is confounded with the process error. User beware: some still do, but those written for environmental applications (e.g., geoR, gstat, FRK) give the correct kriging equations. The difference is most apparent when the kriging variance is computed as $\mathrm{var}(z(\boldsymbol{s}_0)|\boldsymbol{z})$ at location $\boldsymbol{s}_0$, but then interpreted incorrectly as the predictive variance $\mathrm{var}(y(\boldsymbol{s}_0)|\boldsymbol{z})$ of the true process at $\boldsymbol{s}_0$.

2.3. Geophysicists conserve energy but what do data scientists conserve?

Building physical models usually involves ensuring that mass or energy is conserved. If the system is leaking energy, then it needs to be plugged or the energy needs to be followed as it moves into another system.

Now consider a designed experiment where data $\boldsymbol{z}$ are obtained and the statistical model fitted is

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi}\,. \tag{7}$$

The model in (7) is a linear regression with covariate (or design) matrix $\boldsymbol{X}$, $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ consists of random variables that are independent and identically distributed (iid) Gaussian random variables with mean 0 and variance σ^2_ξ.

Suppose that the measuring instrument was carefully calibrated and, in the study protocol, its sample variance from repeated measurements was reported as a number that I shall write as $s^2_\varepsilon > 0$. Note that the uncertainty in the measurements usually needs to be quantified outside the experiment in order to identify the linear-model-error variance. Scientific interest is primarily in $\boldsymbol{\beta}$, but σ^2_ξ is by no means a nuisance parameter. Its restricted maximum likelihood (REML) estimate is:

$$s^2_\xi = (\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-p)\,,$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood (and also the ordinary least squares) estimate of $\boldsymbol{\beta}$.

It could almost be a rule that in any study of this sort, you will see $s^2_\xi > s^2_\varepsilon$. Where did the component of variability, $(s^2_\xi - s^2_\varepsilon)$, go?

Principle 2.3: *In any well defined statistical model, there is conservation of variability.*

Indeed, the model given by (7) is not defined well enough: The error term ξ_i is, in fact, made up of two components of variance:

$$\xi_i = \delta_i + \varepsilon_i\,;\; i = 1, \ldots, n\,,$$

where $\{\varepsilon_i : i = 1, \ldots, n\}$ represent measurement errors. Often forgotten are $\{\delta_i : i = 1, \ldots, n\}$, which represent model errors resulting from using the linear-regression model, $\{\boldsymbol{x}_i^\top\boldsymbol{\beta} : i = 1, \ldots, n\}$, and which are key in accounting for the inexactness of using any model (linear or nonlinear).

The scientific process $\{y_i : i = 1, \ldots, n\}$ is given by $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \delta_i$, and an observation of it is $z_i = y_i + \varepsilon_i$, for $i = 1, \ldots, n$. The HM captures the variability of $\boldsymbol{z}$ and $\boldsymbol{y}$ beautifully,

as follows:

$$\begin{aligned} \boldsymbol{z}|\boldsymbol{y} &\sim \mathrm{N}(\boldsymbol{y}, \sigma_\varepsilon^2 \boldsymbol{I}) \ [\text{or, } \boldsymbol{z} = \boldsymbol{y} + \boldsymbol{\varepsilon}] \,, \\ \boldsymbol{y} &\sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma_\delta^2 \boldsymbol{I}) \ [\text{or, } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\delta}] \,. \end{aligned} \tag{8}$$

Hence, from (8), we obtain the earlier result in vector form:

$$\boldsymbol{z} = (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\delta}) + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{\delta} + \boldsymbol{\varepsilon}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi} \,, \tag{9}$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are independent mean-zero random vectors.

Comparing (7) and (8), we see that the model given by (7) should be augmented with the equation, $\boldsymbol{\xi} = \boldsymbol{\delta} + \boldsymbol{\varepsilon}$, which results in the conservation-of-variability equation, $\sigma_\xi^2 = \sigma_\delta^2 + \sigma_\varepsilon^2$. Consequently, knowing s_ε^2 (an estimate of σ_ε^2 that is obtained outside the experiment) and, having computed the REML estimate of σ_ε^2, an estimate of the model error, σ_δ^2, can be obtained as,

$$s_\delta^2 = s_\xi^2 - s_\varepsilon^2 \tag{10}$$

(provided the right-hand side is non-negative).

In its simplest form, conservation of variability says that the *total variability* is equal to the variability due to *model uncertainty plus* the variability due to *measurement uncertainty*. When variability is captured using a probability space defined by a HM, this principle can be expressed as:

$$\mathrm{var}(\boldsymbol{z}) = \mathrm{var}(\mathrm{E}(\boldsymbol{z}|\boldsymbol{y})) + \mathrm{E}(\mathrm{var}(\boldsymbol{z}|\boldsymbol{y})) \,. \tag{11}$$

For example, consider the HM defined by (8). Since $\mathrm{var}(\boldsymbol{z}) = \mathrm{var}(\boldsymbol{\xi})$ and $\mathrm{var}(\boldsymbol{y}) = \mathrm{var}(\boldsymbol{\delta})$, we have

$$\sigma_\xi^2 \boldsymbol{I} = \sigma_\delta^2 \boldsymbol{I} + \sigma_\varepsilon^2 \boldsymbol{I} \,,$$

and variability is conserved.

This might seem obvious to you, or perhaps even trivial. Again consider (7), and suppose you want to predict an unknown value, y_{n+1}, outside the data set, but only the p-dimensional estimate $\hat{\boldsymbol{\beta}}$ and the covariate, $\boldsymbol{x}_{n+1}$, are at your disposal. Most regression textbooks would say that you should use as predictor, $\boldsymbol{x}_{n+1}^\top \hat{\boldsymbol{\beta}}$. However, a scientist wants to predict the value of the scientific process, $y_{n+1} = \boldsymbol{x}_{n+1}^\top \boldsymbol{\beta} + \delta_{n+1}$. Using the well defined HM (8), its predictor is $\hat{y}_{n+1} = \mathrm{E}(y_{n+1}|\boldsymbol{z}) = \boldsymbol{x}_{n+1}^\top \hat{\boldsymbol{\beta}} + \hat{\delta}_{n+1}$, where $\hat{\delta}_{n+1} = \mathrm{E}(\delta_{n+1}|\boldsymbol{z})$. Since $\{\delta_i\}$ are iid $\mathrm{N}(0, \sigma_\delta^2)$, $\hat{\delta}_{n+1} = \mathrm{E}(\delta_{n+1}|\boldsymbol{z}) = 0$; however, $\mathrm{var}(\delta_{n+1}|\boldsymbol{z})$ is *not zero*, and that has to be recognised in order to perform valid predictive inference, as given below (e.g., Cressie 2020).

The prediction error is $(\hat{y}_{n+1} - y_{n+1})$, and its first two moments are:

$$\begin{aligned}
&\mathrm{E}(\hat{y}_{n+1} - y_{n+1}) = 0\,, \\
&\mathrm{E}(\hat{y}_{n+1} - y_{n+1})^2 = \mathrm{E}(\boldsymbol{x}_{n+1}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2 + \mathrm{E}(\mathrm{var}(\delta_{n+1}|\boldsymbol{z}))\,,
\end{aligned}$$

since the expectation of the cross-product is zero. Thus, if $\mathrm{var}(\delta_{n+1}|\boldsymbol{z})$ were forgotten by the data scientist, such as would happen if the presence of $\{\delta_i\}$ were not recognised in the statistical model, a forecasting decision about future values of the process $\{y_i\}$ would be overly optimistic and potentially harmful.

Principle 2.3 also shows up in the analysis of variance (ANOVA) method, where the 'between sum of squares' corresponds to the variance of the conditional expectation in (11), the 'within sum of squares' corresponds to the expectation of the conditional variance in (11), and the 'total sum of squares' corresponds to the left-hand side of (11). Conservation of variability implicitly includes variability and *covariability*. For example, if two random variables ε_1 and ε_2 have correlation $\rho \neq 0$, then it should be recognised that $\mathrm{var}(\varepsilon_1 + \varepsilon_2) \neq \mathrm{var}(\varepsilon_1) + \mathrm{var}(\varepsilon_2)$ This has often been ignored by scientists doing an error budget (e.g., in remote sensing retrievals; see Connor et al. 2016). Depending on the sign of ρ, $\mathrm{var}(\varepsilon_1) + \mathrm{var}(\varepsilon_2)$, may be either larger than or smaller than the total variability, since $\mathrm{var}(\varepsilon_1 + \varepsilon_2) = \mathrm{var}(\varepsilon_1) + \mathrm{var}(\varepsilon_2) + 2\rho\mathrm{var}(\varepsilon_1)^{1/2}\mathrm{var}(\varepsilon_2)^{1/2}$.

The rules of probability theory can explain easily how the variability can seem to disappear, and then they can show us where to find it. It is less natural to do so simply with empirical distributions, because the pairs $\{(z_i, y_i)\}$ involve the unavailable, hidden variable $\{y_i\}$. Further, if $\{z_i\}$ and $\{x_j\}$ are two sets of observations, there is no guarantee that they will occur in pairs, and hence the empirical correlation, r, might be difficult to obtain.

The famous bias–variance trade-off is another manifestation of Principle 2.3. In the context of estimation of a fixed but unknown parameter θ with an estimator $\hat{\theta}(\boldsymbol{z})$, where recall $\boldsymbol{z}$ is the data vector, the mean-squared error is the sum of the estimator's squared bias and its variance:

$$\mathrm{E}(\hat{\theta}(\boldsymbol{z}) - \theta)^2 = (\mathrm{E}(\hat{\theta}(\boldsymbol{z})) - \theta)^2 + \mathrm{var}(\hat{\theta}(\boldsymbol{z}))\,.$$

Generally speaking, an estimator that decreases the bias will increase the variance, and *vice versa*. The mean-squared error might be decreased using different estimators, but there is a trade-off to be made when trying to decrease both bias and variance at the same time.

2.4. The holy grail: all scales of variation are additive

A statistical analysis cannot get very much simpler than fitting a simple linear regression (a special case of (7)) to $\{(z_i, x_i) : i = 1, \ldots, n\}$, as follows:

$$z_i = \beta_1 + \beta_2 x_i + \xi_i \,;\, i = 1, \ldots, n \,, \tag{12}$$

where $\{\xi_i : i = 1, \ldots, n\}$ are iid $\mathrm{N}(0, \sigma_\xi^2)$. However, is (12) appropriate if $\{z_i : i = 1, \ldots, n\}$ are photon counts at wavelengths in a given band of the electro-magnetic spectrum, or if they are percentages of trace elements in soil?

Exploratory data analysis (EDA) might reveal a highly skewed histogram of $\{z_i\}$. After plotting the histogram of $\{\log z_i\}$, we might achieve a more symmetric appearance; assume for the moment that we do. This would prompt a plot of not only $\{z_i\}$ versus $\{x_i\}$, but also of $\{\log z_i\}$ versus $\{x_i\}$, and the latter might look more linear. Further EDA, in the form of residual plots, after fitting a simple linear regression to the data $\{z_i\}$ and one to the transformed $\{\log z_i\}$, would be carried out. And, for example, it might be that the residual variability of $\{z_i\}$ appears to increase with x (i.e., heteroskedasticity), but the residual variability of $\{\log z_i\}$ shows no dependence on x (i.e., homoskedasticity).

Admittedly, this is a story, not methodology, but in my experience with analysing data, it happens enough to propose the next principle (Cressie 1985).

Principle 2.4: *Seek a transformation of the scientific process where all components of variation act and interact additively.*

This principle looks more like a doctrine but, as I explain below, a fitting of (12) to the data $\{z_i\}$ is following this principle, where the transformation is simply the identity.

Suppose the quantity y has variabilities that can be expressed as 'large-scale' and 'small-scale.' Scientists usually put their theories in the large scale and their errors (both in the theory and in the measurements) in the small scale. Statistical scientists know that to make inference on the large scale of the scientific process, its small scale has to be modelled well, usually by a random error term, δ. However, the model I now give for the scientific process is more basic, given in terms of those large and small scales of variability: It is additive and of the following form,

$$y = (\mu^{(1)} + \ldots + \mu^{(p)}) + (\delta^{(1)} + \ldots + \delta^{(N)}) \,, \tag{13}$$

where p and N are positive integers and, for convenience, the subscript 'i' has been dropped. For example, the large-scale variation in simple linear regression has $p = 2$, $\mu^{(1)} = \beta_1$, and $\mu^{(2)} = \beta_2 x$. For the small-scale variation in (13), N is large, and $\{\delta^{(l)} : l = 1, \ldots, N\}$ are physically interpretable components with variabilities that are very small but such that their total variability is the variability of the difference, $y - \mu^{(1)} - \ldots - \mu^{(p)}$; see Principle 2.3.

The model (13) is a physical model where the large scales act additively with each other, the small scales act additively with each other, and the large scales and the small scales interact additively. It can be made statistical by assuming that these many small-scale effects, $\delta^{(1)}, \ldots, \delta^{(N)}$, are random, statistically independent, have mean zero and variances that are $O(1/N)$, with the same leading coefficient, denoted here as σ^2_δ. Now let N be large and define the small-scale variations collectively as δ; then (13) can be written as,

$$y = \mu^{(1)} + \ldots + \mu^{(p)} + \delta\,,$$

where

1. $\mathrm{E}(y) = \mu^{(1)} + \ldots + \mu^{(p)}$ (additivity),
2. $\mathrm{var}(y) = \sigma^2_\delta$ (homoskedasticity),
3. $\delta \sim \mathrm{N}(0, \sigma^2_\delta)$ (central limit theorem).

The regression (7) is obtained when the large-scale effects $\{\mu^{(k)} : k = 1, \ldots, p\}$ are identified with $\{\beta_k x_k : k = 1, \ldots, p\}$. It is the default model used in much of science, but this discussion shows that it originates from the imposition of additivity within and between scales of variability. From Principle 2.2, the measurement of a phenomenon should be separated from how the phenomenon behaves in nature and, indeed, the generalised linear model does this through a link function (McCullagh & Nelder 1989). Hence, Principle 2.4 covers many statistical models and offers a plausible explanation of why a linear relation, homoskedasticity, and Gaussianity are often found to occur together after a transformation of the data (e.g., Cressie 1978). Of course, it is easy to construct probability models where the principle does not hold, but one should keep in mind that probability theory is used to model nature's variability, not the other way around.

When nonlinearity is inherent to the physical system, such as would occur when there are barriers or thresholds, the quest looks to be futile. However, after following Principle 2.2, it may just be that the grail, Principle 2.4, is hidden deep in the process model (Berliner, Wikle & Cressie 2000).

2.5. Could swans be red?

Taleb (2007) published a book about 'Black Swan' events that, when they happen, are considered to have been unpredictable. Such events could have a major impact on Earth's environment or, as has happened in 2020, on global-population health and economies around the globe.

This black-swan meme has its roots in seventeenth-century Europe. Up to then, no swans had ever been observed that were black but, in 1697, Dutch explorers became the

first Europeans to observe black swans during a voyage that took them to Western Australia. The implication in Taleb's book is that if the scientific world in the 1600s could not predict black swans, how can scientists predict catastrophic environmental events in the twenty-first century?

My response is that 300+ years ago the best scientific minds in Europe were too certain about their science; that is, if they had been asked to put probabilities (according to abundance) on the colours that swans could be: red, orange, . . ., violet, white, black, they would have put $0, 0, \ldots, 1, 0$, which is a well defined probability mass function. Certainly, black-swan events are *not* predictable if the scientific model is 100% certain that they do not exist. Because their probability model gave black swans and indeed coloured (including red) swans, zero probability, this unimagined event did not emerge out of the 'ether' in subsequent inferences, until one was observed. Do we have to wait until a highly unusual event occurs before we are forced to change the probability model? The real lesson from the black-swan meme is that scientific knowledge is never perfect, that modellers need to explore the parameter space thoroughly, and that they need to 'spend' some probability on highly unusual events.

At the time of writing this article, our species is under attack from a virus that was originally called 'novel coronavirus,' so new that it took several weeks before the infection it caused had a name: COVID-19. Virologists certainly had assigned a small positive probability that each new decade would have its own 'novel' pandemic. However, politicians (and most economists) appear to have put zero probability on a severe, worldwide economic disruption.

Given a swine-flu-like pandemic has occurred, the conditional probability of some economic disruption is *not zero*. But, given a severe pandemic has occurred, the conditional probability of severe economic disruption is substantial. This conditional probability is then multiplied by the probability of a severe pandemic, which is by no means zero given the ability of viruses to mutate and occasionally jump from animals to humans. The product of these two is the joint probability of a severe pandemic followed by severe economic disruption, which is not negligible. This has happened twice in the last hundred years, and it will likely happen more often with humans and animals sharing in an ever-more-crowded environment.

Principle 2.5: *When building probability models, look carefully where zero probabilities are assumed (perhaps implicitly) and, with the same care, move appropriate probabilities away from zero. Calculate joint probabilities from products of conditional (not marginal) probabilities, unless entropy is maximal.*

Unfortunately, uncertainty quantification through joint probabilities all too often comes from multiplying marginal probabilities as if each event in the causal chain were independent. It does not 'hurt' for small probabilities to be assigned to Pr(bird is colour $c \mid$ bird is a swan)

where c also covers the colour 'black.' Then Pr(bird is a swan and it is black) = Pr(bird is black | bird is a swan) $\times$ Pr(bird is a swan). The point being made here is philosophical, and it could be applied to Pr(pandemic followed by global economic disruption), for different severities of both events. The worst thing would be to make Pr(global economic disruption | pandemic) equal to zero, since then there would be a lack of planning for the health and economic crises brought on by a pandemic (Mackenzie 2020).

If nothing were known about the conditional probability, a fall-back is the maximum-entropy model where the marginal probability, Pr(bird is colour c), is used in place of Pr(bird is colour c | bird is a swan). The marginal probability is not zero, since it is based on abundance and, for example, there are birds that are predominantly coloured red (e.g., Australia's king parrot). The maximum-entropy principle is discussed in Cressie, Richardson & Jaussent (2004).

Experiences over the last 100 years mean that Pr(global economic disruption), whether caused by a pandemic or not, should not be zero, yet governments described the events of March–April 2020 as unimaginable. Events of small (but not zero) probabilities with consequences expressed through a loss function, allow a non-zero expected loss to be calculated, which can then be used to make optimal decisions that mitigate the consequences (e.g., see Berger 1985 for a discussion of decision theory).

I have just presented five statistical principles that should be useful in Data Science. Data often come with location information, in which case the data scientist will likely be using spatial-analysis methods. In the next section, I present three principles that are specific to Spatial Statistics.

3. A few spatial statistical principles

Those of us who work in Spatial Statistics will know of Tobler's First Law of Geography (Tobler 1970). In spatial statistical science, it really is a 'principle' rather than a 'law' and, in the following subsections, I shall present it and two other principles that I have found useful in this area of research.

3.1. Patches in close proximity are commonly more alike ...

In his famous 1935 book on experimental design, Fisher (1935, p. 66), wrote: 'After choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart.' A spatial statistician sees this as the making of a principle but, at that time, Fisher made a sharp right turn. In his analyses of field trials he applied randomisation to eradicate the pest: spatial correlation! Cressie & Wikle (2011, Ch.1 and Ch. 4) give

some historical perspective to the work of researchers who took roads less travelled and developed this vibrant area we now call Spatial Statistics. Some of this development comes from Geography, and so it is fitting that the first and most important principle in this section has become known as the First Law of Geography. Originally articulated by Tobler (1970), it is given here in exactly his words.

Principle 3.1 *Everything is related to everything else, but near things are more related than distant things (Tobler 1970).*

This principle is at the core of what we do in spatial and spatio-temporal statistics. For example, in remote sensing of Earth's surface, the scene of interest D contains *many* retrievals, $\{z(\boldsymbol{s}_i) : i = 1, \ldots, n\}$, where $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ are the (lon, lat) locations of the n data inside D (retrieved over a short time period). There is a hidden spatial process $\{y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ that the geophysicist would like to infer and, in a spatial HM, a spatial statistical model for $\{y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is built around Principle 3.1.

For example, consider a simple process model, appropriate for a small scene D:

$$y(\boldsymbol{s}) = \beta_1 + \beta_2 \text{lat}(\boldsymbol{s}) + \delta(\boldsymbol{s}), \text{ for } \boldsymbol{s} \in D,$$

where the mean of the process is a linear function of latitude, and $\{\delta(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is a spatial error process with mean zero and stationary covariance function, $C_\delta(\boldsymbol{h}) = \text{cov}(\delta(\boldsymbol{s}+\boldsymbol{h}), \delta(\boldsymbol{s})) = \sigma^2_\delta \cdot \exp(-\|\boldsymbol{h}\|/\phi)$. Notice that $C_y(\boldsymbol{h}) = \text{cov}(y(\boldsymbol{s}+\boldsymbol{h}), y(\boldsymbol{s})) = C_\delta(\boldsymbol{h})$; $C_y(\boldsymbol{0}) = C_\delta(\boldsymbol{0}) = \sigma^2_\delta$; and the scale parameter ϕ controls how 'more related' things are. With this parameterisation, $\phi = 0$ is the degenerate case of no spatial correlation and, as ϕ increases, the spatial correlation increases for a given distance $\|\boldsymbol{h}\|$. The data model in this example is also simple, that of additive independent measurement error. The data vector is $\boldsymbol{z} = (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n))^\top$ and

$$z(\boldsymbol{s}_i) = y(\boldsymbol{s}_i) + \varepsilon_i\,;\; i = 1, \ldots, n,$$

where $\{\varepsilon_i : i = 1, \ldots, n\}$ are independent mean-zero errors.

Assuming for the moment that all the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2_\delta, \phi, \sigma^2_\varepsilon)^\top$ are known, inference on $\{y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ comes from summaries of the predictive distribution, $[\{y(\boldsymbol{s}) : \boldsymbol{s} \in D\}|\boldsymbol{z}]$. For a Gaussian process model $y(\cdot)$ and Gaussian measurement errors $\{\varepsilon_i\}$, the predictive distribution is also a Gaussian process whose first two moments, $\text{E}(y(\boldsymbol{s})|\boldsymbol{z})$ and $\text{cov}(y(\boldsymbol{s}), y(\boldsymbol{u})|\boldsymbol{z})$, for $\boldsymbol{s}, \boldsymbol{u} \in D$, can be obtained analytically. From Section 2.2, a common predictor of $y(\boldsymbol{s})$ is $\hat{y}(\boldsymbol{s}) = \text{E}(y(\boldsymbol{s})|\boldsymbol{z})$, whose statistical properties are needed for inference. It is straightforward to see that $\text{E}(\hat{y}(\boldsymbol{s})) = \text{E}(y(\boldsymbol{s}))$ (i.e., the predictor is

unbiased) and the mean-squared predictor error is:

$$\mathrm{E}(\hat{y}(\boldsymbol{s}) - y(\boldsymbol{s}))^2 = \mathrm{E}(\mathrm{var}(y(\boldsymbol{s})|\boldsymbol{z})) = \mathrm{var}(y(\boldsymbol{s})|\boldsymbol{z})\,,$$

which is the predictive variance (the last equality being due to the Gaussian assumptions made). Further summaries might come from well chosen percentiles (e.g., 2.5%, 25%, 50%, 75%, 97.5%) of $[y(\boldsymbol{s})|\boldsymbol{z}]$.

3.2. What is one person's mean function could be another person's spatial error

In spatial statistics, it might appear that the spatial-prediction problem is quite difficult, because the number of predictions to be made is often greater than the number of data. Even for the problem of parameter estimation, the number of 'degrees of freedom' in n spatially dependent data will not be n, as I now show. Applying Principle 3.1 with a stationary spatial covariance model $C_z(\boldsymbol{h}) = \mathrm{cov}(z(\boldsymbol{s}+\boldsymbol{h}), z(\boldsymbol{s}))$, for $\boldsymbol{h} \in \mathbb{R}^d$, that exhibits only positive correlations, it is easy to see that for $\overline{z} = (1/n)\sum_{i=1}^n z(\boldsymbol{s}_i)$,

$$\mathrm{var}(\overline{z}) = (1/n)\{\sigma_z^2 + 2\sum_{i<j}\sum C_z(\boldsymbol{s}_i - \boldsymbol{s}_j)/n\} > \frac{\sigma_z^2}{n}\,,$$

where $\sigma_z^2 = C_z(\boldsymbol{0})$. Hence one can define the *effective degrees of freedom*, n_{eff}, as

$$n_{\mathrm{eff}} = \sigma_z^2/\mathrm{var}(\overline{z}) < n\,;$$

that is, under the almost ubiquitous spatial model of positive spatial correlation, $n_{\mathrm{eff}} < n$. In the remote sensing context where there are many observations taken over a short period of time, Zhang, Cressie & Wunch (2017) gave an example where $n = 2961$ and $n_{\mathrm{eff}} = 202.3$, less than one-tenth of the sample size!

These calculations are a warning that, in the spatial (and spatio-temporal) setting, intuition learned from the 'iid errors' model has to be modified, sometimes substantially. Critically, the probability model that captures the spatial variability has to be well specified, and we now discuss how this can be done.

The classical spatial statistical model consists of a deterministic mean process, $\{\mu(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, and a random spatial error process, $\delta(\boldsymbol{s}) = y(\boldsymbol{s}) - \mu(\boldsymbol{s})$, so that

$$y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \delta(\boldsymbol{s})\,,\ \text{for}\ \boldsymbol{s} \in D\,; \tag{14}$$

see Cressie (1993, Ch. 2). It is often a personal choice what components go into $\mu(\boldsymbol{s})$. For linear regression, $\mu(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^\top\boldsymbol{\beta}$, for $\boldsymbol{s} \in D$, but the set of possible covariates, $\boldsymbol{x}(\boldsymbol{s})$,

typically comes from the modeller's subject-matter knowledge, augmented by *spatial trend* terms that are linear and higher-order functions of $\boldsymbol{s} = (s_1, \ldots, s_d)^\top$.

What is the effect of not having included an important spatially varying covariate, $x_{p+1}(\boldsymbol{s})$ say, in $\mu(\boldsymbol{s})$? From Principle 2.3, $\{\delta(\boldsymbol{s})\}$ has to absorb the contribution to spatial variability that $\{x_{p+1}(\boldsymbol{s})\}$ would have made had it been included in the regression. This means that a fixed effect that has been inadvertently left out will be picked up in the spatial statistical model (14), as a random effect.

Principle 3.2: *Assume a decomposition of a spatial process into fixed effects plus random effects. While it is not unique, the decomposition must be chosen to conserve variability.*

This principle is a refinement of Principle 2.3 that is adapted here for Spatial Statistics. The variability of the deterministic-mean component is measured differently from that of a random-error component. Define the *regional variance* of $\mu(\cdot)$ as,

$$s_\mu^2 = \left(\frac{1}{|D|}\right) \int_D (\mu(\boldsymbol{s}) - \overline{\mu})^2 \, \mathrm{d}\boldsymbol{s} \,,$$

where $\overline{\mu} = (1/|D|) \int_D \mu(\boldsymbol{s}) \, \mathrm{d}\boldsymbol{s}$ is the regional average of $\mu(\cdot)$ (Lahiri et al. 1999). Suppose that a stationary covariance function, $C_\delta(\cdot)$, describes the covariability in $\{\delta(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, with $\sigma_\delta^2 = C_\delta(\mathbf{0})$; recall that $C_\delta(\boldsymbol{h}) = \mathrm{cov}(\delta(\boldsymbol{s}+\boldsymbol{h}), \delta(\boldsymbol{s}))$, for $\boldsymbol{h} \in \mathbb{R}^d$. Let $\hat{C}_\delta(\cdot)$ denote the empirical covariance function, and $s_\delta^2 = \hat{C}_\delta(\mathbf{0})$. Then, according to Principle 3.2, *approximately* speaking, $s_\mu^2 + s_\delta^2$ should not change as different decompositions given by (14) are fitted. Now suppose that the additional covariate vector, $\boldsymbol{x}_{p+1} = (x_{p+1}(\boldsymbol{s}_1), \ldots, x_{p+1}(\boldsymbol{s}_n))^\top$, is included in the regression, and it is not in the column space of $\boldsymbol{X}$. Then generally, the new empirical covariance function $\hat{C}_\delta(\cdot)$, yields a new estimate $s_\delta^2 = \hat{C}_\delta(0)$. The principle says that this new s_δ^2 should decrease to allow for the additional spatial variability in $\{\mu(\boldsymbol{s}) : \boldsymbol{s} \in D\}$. (Understandably, the general shape of the new empirical covariance function will change as well.)

How will we ever know which model is better after each is fitted to the same spatial data $\{z(\boldsymbol{s}_i)\}$? Model-selection criteria such as 'Akaike Information Criterion,' 'Deviance Information Criterion,' and cross-validation will remove bad models but, in the 'difficult middle,' there is no way to know whether a graph of $\{z(\boldsymbol{s}_i)\}$ versus $\{x_{p+1}(\boldsymbol{s}_i)\}$ is showing the behaviour of a component of the mean function or the behaviour of a component of the error process. Germane to Principle 3.2, it is well known to spatial statisticians that a single simulation of a strongly spatially dependent random effect, $\{\delta(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, can look like deterministic spatial trend. If a replicate of the data were available, and if the analogous plot showed the same behaviour as seen in the original plot, then $\{x_{p+1}(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ would probably belong in the mean function $\mu(\cdot)$. From just one realisation, the decomposition (14)

is not unique, and hence what is one person's mean function could be another person's spatial error.

3.3. COS is the DNA of Spatial Statistics

Most of us have heard of, or encountered, *Simpson's Paradox*, which basically says that two variables x and y could be positively correlated but, *conditional* on a third variable w, it is perfectly acceptable that x and y be negatively correlated (or uncorrelated, or positively correlated)! Usually, Simpson's Paradox is expressed in terms of categorical data, which we could do here by defining ordinal categories, or bins, from the ranges of w, x, and y.

Let a sample $\{(w_i, x_i, y_i) : i = 1, \ldots, n\}$ be assigned to the pre-defined bins, creating a three-way contingency table with counts in each cell of the table. A two-way table showing (marginal) dependence between binned x and binned y is obtained by aggregating the counts across the third variable, namely binned w. Then a measure of dependence in the two-way table (e.g., the statistic *gamma*, due to Goodman & Kruskal 1954) could show positive dependence, but the same measure applied to the two-way tables of binned x and binned y *conditional* on each of the values of binned w, could all show negative (or no, or positive) dependence.

What can you do about it? First, understand why it happens, and then spend the rest of your data-science career looking for those lurking variables like w that you may have missed! It manifests in other settings as well, as I now discuss.

Let x and y be jointly Gaussian random variables and related through the simple linear regression model given by (12); the probability model of how y varies with x is as follows: Conditional on x, the random variable y is Gaussian with mean and variance, respectively,

$$\begin{aligned} \mathrm{E}(y|x) &= \mathrm{E}(y) + \{\rho_{xy}\mathrm{var}(y)^{1/2}/\mathrm{var}(x)^{1/2}\}\{x - \mathrm{E}(x)\}\,, \\ \mathrm{var}(y|x) &= \mathrm{var}(y)\{1 - \rho_{xy}^2\}\,, \end{aligned}$$

where $\rho_{xy} = \mathrm{corr}(x, y)$.

Now consider regression of y on both x and w, where again joint Gaussianity of random variables w, x, and y is assumed. To make the calculations easier to follow, in the rest of this subsection consider the simple case where $\mathrm{E}(w) = \mathrm{E}(x) = \mathrm{E}(y) = 0$, and $\mathrm{var}(w) = \mathrm{var}(x) = \mathrm{var}(y) = 1$. Then the covariances, $\mathrm{cov}(x, y)$, $\mathrm{cov}(w, y)$, and $\mathrm{cov}(x, w)$ are identical to correlations, which are denoted here as ρ_{xy}, ρ_{wy}, and ρ_{xw}, respectively. The regression lines to be compared are,

$$\mathrm{E}(y|x, w) = \left\{\frac{\rho_{wy} - \rho_{xy}\rho_{xw}}{1 - \rho_{xw}^2}\right\} w + \left\{\frac{\rho_{xy} - \rho_{wy}\rho_{xw}}{1 - \rho_{xw}^2}\right\} x\,, \tag{15}$$

and

$$\mathrm{E}(y|x) = 0 + \rho_{xy} \cdot x\,.$$

Then, conditional on w,

$$\begin{aligned}\mathrm{corr}(x, y|w) &= \left\{ \frac{\rho_{xy} - \rho_{wy}\rho_{xw}}{1 - \rho_{xw}^2} \right\} \frac{\mathrm{var}(x|w)^{1/2}}{\mathrm{var}(y|w)^{1/2}} \\ &= \left\{ \frac{\rho_{xy} - \rho_{wy}\rho_{xw}}{(1 - \rho_{xw}^2)^{1/2}(1 - \rho_{wy}^2)^{1/2}} \right\}, \end{aligned} \tag{16}$$

which is to be compared to $\mathrm{corr}(x, y) = \rho_{xy}$.

If w has zero correlation with both x and y, then from (16), $\mathrm{corr}(x, y|w) = \mathrm{corr}(x, y) = \rho_{xy}$, which is an intuitively reasonable result. In general, $\rho_{wy} \neq 0$ and $\rho_{xw} \neq 0$, so from (15) it is clear that a lurking variable w can wreak havoc on any honest attempt to interpret a simple linear regression of y on x. But, what does Simpson's Paradox have to do with spatial statistics?

The answer is 'plenty,' if you think about w as being a variable that describes a range of geographical strata. For example, the Australian Bureau of Statistics divides Australia up into *small areas*, at different levels of aggregation. A study of y =mean weekly income in NSW, Australia, regressed on selected demographic variables $\boldsymbol{x}$ at the finest level of aggregation, was given in Burden, Cressie & Steel (2015). Given the great divide in Australia between 'city' and 'country,' the most obvious variable w to choose is: $w = 1$ if the small area is in Greater Sydney (city), and otherwise $w = 0$ (country). In this set-up, it is easy to imagine that the two regressions, $\mathrm{E}(y|\boldsymbol{x}, w = 1)$ and $\mathrm{E}(y|\boldsymbol{x}, w = 0)$, would result in more interpretable results than the single regression, $\mathrm{E}(y|\boldsymbol{x})$. (In fact, Burden, Cressie & Steel 2015, used Principle 3.2 and captured the 'geography' by using a spatial error term rather than a geographic covariate w.)

Simpson's Paradox is potentially everywhere in the spatial context, because regressions of y on x can be done at different levels of spatial aggregation. A regression of y on x at the finest level of aggregation may show positive dependence, but when the variables are aggregated to a coarser level, the regression may show negative (or no, or positive) dependence! In Sociology, this phenomenon has been called the *ecological effect* (Robinson 1950), an unfortunate and misleading name that has no direct connection to Ecology. In Geography, the combination of Simpson's Paradox and the ecological effect has been called the Modifiable Areal Unit Problem, or MAUP; a spatial statistical discussion of MAUP is given in Cressie (1996).

Because it is so natural to aggregate processes over space, meta-data in the spatial data set should include the *spatial support* of each datum. Define $y(B) = (1/|B|) \int_B y(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}$;

then $y(B)$ has *spatial support* B with volume, $|B| = \int_B \mathrm{d}\boldsymbol{s} > 0$. In geostatistics, B is called a *block*, and we say that $y(B)$ is an aggregation (or block average) of the original process. Then a *change-of-support* (COS) is said to have occurred from point support to spatial support B, with a corresponding change of statistical properties. Data on mutually exclusive supports, $\{B_1, \ldots, B_n\}$, are typically represented as:

$$z(B_i) = y(B_i) + \varepsilon_i\,;\; i = 1, \ldots, n\,.$$

The simplest model would be $\{\varepsilon_i : i = 1, \ldots, n\}$ iid $\mathrm{N}(0, \sigma^2_\varepsilon)$, but modifications to account for the protocols of measurement and the possible overlap of the supports $\{B_i\}$ have been developed (e.g., Wikle & Berliner 2005).

The two earlier principles of Spatial Statistics (nearby things are more alike; and decompose spatial variability into fixed effects plus random effects) are joined here by a change-of-support principle.

Principle 3.3: *The variance of the aggregation, $y(B)$, is a decreasing function of the volume, $|B|$.*

This is the most important of a number of COS properties (e.g., Gotway & Young 2002), whose discussion I have left for a future time.

Let $y(\boldsymbol{s}) = y_0 + \delta(\boldsymbol{s})$, for $\boldsymbol{s} \in D$, where y_0 is a non-degenerate random variable with possibly non-zero mean and independent of $\{\delta(\boldsymbol{s}) : \boldsymbol{s} \in D\}$. Then for $B \subset D$,

$$\mathrm{var}(y(B))) = \sigma^2_0 + \mathrm{var}(\delta(B))\,, \tag{17}$$

where $\sigma^2_0 = \mathrm{var}(y_0) > 0$. Therefore, if $\mathrm{var}(\delta(B))$ decreases to 0 as the volume $|B|$ increases, $\mathrm{var}(y(B))$ decreases to $\sigma^2_0 > 0$. It is this type of behaviour that is of interest to geoscientists analysing remote sensing data. In that literature (reviewed in Cressie 2018), they distinguish between two types of error, 'systematic error' and 'random error,' as follows.

Random error: An error is a random error if the average of a collection of n of them has variability that decreases to zero like $1/n$, as n becomes large.

Systematic error: An error is a systematic error if the average of a collection of n of them has variability that does not decrease to zero, as n becomes large.

These might be considered 'verbal working definitions,' but a statistical scientist looks at these and tries to find an appropriate probability model. The ones I give below are building blocks for parts of a HM. In practice, the most difficult aspect is to know which errors are of which type and how to group them together for inference.

For 'random error,' the obvious probability model is: Let $\delta(\boldsymbol{s}_1), \ldots, \delta(\boldsymbol{s}_n)$ be iid random variables with mean 0 and variance σ^2_δ. Now the average, $\overline{\delta} = (\sum_{i=1}^n \delta(\boldsymbol{s}_i))/n$, has

$\mathrm{E}(\overline{\delta}) = 0$ and $\mathrm{var}(\overline{\delta}) = \sigma_\delta^2/n$, which decreases to zero like $1/n$. In a 'data-rich' situation and with enough averaging of the data, random errors can be shown to be annihilated by the averaging (using the law of large numbers). In a spatial setting, Principle 3.1 suggests that the independence assumption between the $\{\delta(\boldsymbol{s}_i)\}$ is not appropriate. Under increasing-domain asymptotics, it can be shown that $\mathrm{var}(\overline{\delta})$ still decreases like $1/n$ (Cressie 1993), however strong spatial dependence can make the errors look more systematic than random (Morris & Ebey 1984).

For 'systematic error,' an obvious probability model is a random-effects model: Let $e(\boldsymbol{s}_1), \ldots, e(\boldsymbol{s}_n)$ be written as

$$e(\boldsymbol{s}_i) = y_0 + \delta(\boldsymbol{s}_i)\,; \quad i = 1, \ldots, n\,,$$

where y_0 has mean 0 and variance $\sigma_0^2 > 0$, and y_0 and $\{\delta(\boldsymbol{s}_i)\}$ are independent. Now, if $\{\delta(\boldsymbol{s}_i)\}$ are iid mean 0 and variance σ_δ^2, then the average error $\overline{e} = \sum_{i=1}^n e(s_i)/n$ has variance,

$$\mathrm{var}(\overline{e}) = \sigma_0^2 + \sigma_\delta^2/n\,.$$

This does not decrease to zero as n becomes large, and hence $e(\cdot)$ is a form of systematic error. Zhang, Cressie & Wunch (2019) used spatial dependence in the $\{\delta(\boldsymbol{s}_i)\}$ and an additive random effect y_0 with mean 0, as part of the OCO-2 calibration of remote sensing data to 'ground-truth' data.

Consider spatial prediction errors, $\{\hat{y}(\boldsymbol{s}_i) - y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$, which are located at $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ in the spatial domain D. Typically, these prediction errors have individual means that are not zero, and it is often the case in Spatial Statistics that there is no replication to estimate them. A way out of the conundrum is to assume exchangeability (e.g., Spiegelhalter, Abrams & Myles 2004, Section 3.17), which results in a spatial statistical model that exhibits both systematic error and random error; see Cressie (2018, Rejoinder).

4. Epilogue

In the previous sections, I presented five principles for Statistical Science and three special ones for Spatial Statistics. In what follows, I add a ninth principle that speaks to all of Science, not just Data Science, and I give some concluding remarks.

4.1. You can't add apples and oranges

There is one further principle that has been my 'rock' in the diverse applied-statistics projects in which I have participated. I once saw a cartoon that showed a small town's 'Welcome' sign, and it looked something like this:

Centerville

Population	:	2,390
Elevation	:	862 feet
Total	:	3,252

Keeping track of units is a fundamental part of all of science, including Data Science.

Principle 4.1: *Only add quantities that have the same units. Multiply quantities that have different units, and cancel units from the product whenever possible. Take logs, exponents, and other special functions of unit-free quantities.*

The first part of the principle is illustrated with the cartoon I referred to above. You might say you would never add apples and oranges but, if in (7), z is measured in petagrams (Pg) of carbon in Earth's atmosphere and $x_1 = 1$, $x_2 = t$ (in years), and $x_3 = t^2$, then a unit analysis using the second part of the principle would reveal that regression coefficient β_1 is in units of Pg, β_2 is in Pg/yr, and β_3 is in Pg/yr^2. Principle 4.1 is respected, since it is the regression components $\{\beta_k x_k\}$ that are added. However, it is the regression coefficients $\{\beta_k\}$ that are often interpreted, so I often standardise the covariates by, respectively, subtracting their averages and dividing by their sample standard deviations. Then, after this standardisation, $\beta_1, \beta_2, \ldots, \beta_k$ all have the same units as those of z.

Statistical scientists know that probabilities have no units. Using Principle 4.1, a unit analysis of the fundamental probability equation, $\int f(y)\, dy = 1$, where $f(\cdot)$ is the probability density function of the random variable y, whose units are, say, Pg of carbon, reveals that $f(y)$ has units of (Pg)$^{-1}$. Using the same principle, $\mathrm{E}(y)$ has units of Pg, and so forth. While probability density functions have units, cumulative distribution functions and probability mass functions do not.

The last part of Principle 4.1 applies to any special function that can be expressed as a Taylor series. The most common ones in science are logs and exponents, whose Taylor series are

$$\log_e(1 + x) = x - x^2/2 + x^3/3 - \ldots$$

and

$$e^x = 1 + x + x^2/2! + x^3/3! + \ldots,$$

which only make sense when x has no units; see the first part of Principle 4.1. Euler's number, e, has no units, because $e = \lim_{n\to\infty}(1 + 1/n)^n$.

Counts, percentages, and correlations have no units. Also, the Box-Cox transformation (Box & Cox 1964),

$$g_\lambda(x) = \begin{cases} (x^\lambda - 1)/\lambda\,, & \text{for } \lambda \in (-\infty, 0) \cup (0, \infty) \\ \log x\,, & \text{for } \lambda = 0 \end{cases},$$

only makes sense if x has been rendered unit-free. One way to accomplish this is to specify x relative to a given standard.

Every term in the process model (and data model) in a HM, should be assigned their rightful units. The scientific models that make up parts of the process model are sometimes derived theoretically, sometimes empirically. Beware of a 'scientific constant.' It may be an (estimated) regression coefficient, in which case its units (the ratio of the dependent variable's units to the corresponding covariate's units) are key, since that 'constant' might change if the units change.

In conclusion, it is a critical part of every collaboration to do a 'unit analysis,' which avoids obvious mistakes made by confusing different parts of different systems of units (e.g., metric and imperial), as well as the more subtle ones discussed above.

4.2. These are my nine principles

Principle 2.1: *Establish a true model (TM), perhaps different from the scientist's working model (WM). Critically, compute the TM-distributional properties of the WM estimators.* [All models are wrong . . . but some are wrong-er than others.]

Principle 2.2: *Build statistical models conditionally, through a data model and a process model. Infer the unknown process from the predictive distribution.* [What you see is not what you want to get.]

Principle 2.3: *In any well defined statistical model, there is conservation of variability.* [Geophysicists conserve energy but what do data scientists conserve?]

Principle 2.4: *Seek a transformation of the scientific process where all components of variation act and interact additively.* [The holy grail: all scales of variation are additive.]

Principle 2.5: *When building probability models, look carefully where zero probabilities are assumed (perhaps implicitly) and, with the same care, move appropriate probabilities away from zero. Calculate joint probabilities from products of conditional (not marginal) probabilities, unless entropy is maximal.* [Could swans be red?]

Principle 3.1 *Everything is related to everything else, but near things are more related than distant things (Tobler 1970).* [Patches in close proximity are commonly more alike . . .]

Principle 3.2: *Assume a decomposition of a spatial process into fixed effects plus random effects. While it is not unique, the decomposition must be chosen to conserve variability.* [What is one person's mean function could be another person's spatial error.]

Principle 3.3: *The variance of the aggregation, $y(B)$, is a decreasing function of the volume, $|B|$.* [COS is the DNA of Spatial Statistics.]

Principle 4.1: *Only add quantities that have the same units. Multiply quantities that have different units, and cancel units from the product whenever possible. Take logs, exponents, and other special functions of unit-free quantities.* [You can't add apples and oranges.]

4.3. A disclaimer

These nine principles of Statistical Science are personal, leading to a certain amount of self-referencing, but I hope others find them useful. There are no theorem-proof results, but there are back-of-the-envelope calculations that I use to justify the principles in simple settings. At the very least, they should give data scientists boundaries in their analytics that should be respected, and criteria by which supervised and unsupervised machine-learning methods could be assessed.

4.4. Happy birthday!

I once introduced Adrian Baddeley at a conference session I was chairing (ASC2010, Fremantle, Western Australia) as a national treasure. I repeat it here, and I wish him a very happy 65th birthday!

References

BERGER, J.O. (1985). *Statistical Decision Theory*. Springer, New York, NY.

BERLINER, L.M., WIKLE, C.K. & CRESSIE, N. (2000). Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of Climate* **13**, 3953–3968.

BOX, G.E.P. & COX, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26**, 211–252.

BURDEN, S., CRESSIE, N. & STEEL, D. (2015). The SAR model for very large datasets: A reduced-rank approach. *Econometrics* **3**, 317–338.

CONNOR, B., BÖSCH, H., MCDUFFIE, J., TAYLOR, T., FU, D., FRANKENBERG, C., O'DELL, C., PAYNE, V.H., GUNSON, M., POLLOCK, R., HOBBS, J., OYAFUSO, F. & JIANG, Y. (2016). Quantification of uncertainties in OCO-2 measurements of XCO2: Simulations and linear error analysis. *Atmospheric Measurement Techniques* **9**, 5227–5238.

CRESSIE, N. (1978). Removing nonadditivity from two-way tables with one observation per cell. *Biometrics* **34**, 505–513.

CRESSIE, N. (1985). When are relative variograms useful in geostatistics? *Journal of the International Association for Mathematical Geology (*now *Mathematical Geosciences)* **17**, 693–702.

CRESSIE, N. (1993). *Statistics for Spatial Data, rev. edn.* Wiley, New York, NY.

CRESSIE, N. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems* **3**, 159–180.

CRESSIE, N. (2018). Mission CO_2ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide (with discussion and a rejoinder). *Journal of the American Statistical Association* **113**, 152–181.

CRESSIE, N. (2020). When is it Data Science and when is it Data Engineering? Comment on "Prediction, Estimation, and Attribution" by B. Efron. *Journal of the American Statistical Association,* **115**, 660–662.

CRESSIE, N., RICHARDSON, S. & JAUSSENT, I. (2004). Ecological bias: Use of maximum-entropy approximations. *Australian and New Zealand Journal of Statistics* **46**, 233–255.

CRESSIE, N. & WIKLE, C.K. (2011). *Statistics for Spatio-Temporal Data.* John Wiley & Sons, Hoboken, NJ.

FISHER, R.A. (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh, UK.

GOODMAN, L.A. & KRUSKAL, W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* **49**, 732–764.

GOTWAY, C. & YOUNG, L.J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association* **97**, 632–648.

LAHIRI, S.N., KAISER, M.S., CRESSIE, N. & HSU, N.J. (1999). Prediction of spatial cumulative distribution functions using subsampling (with discussion and a rejoinder). *Journal of the American Statistical Society* **94**, 86–110.

MACKENZIE, D. (2020). *Covid-19: The Pandemic that Should Never Have Happened, and How to Stop the Next One.* The Bridge Street Press, London, UK.

MATHERON, G. (1963). *Traité de Geostatistique Appliquée, Tome II: Le Krigeage.* Memoires du Bureau de Recherche Géologique et Minière, No. **24**. Paris, France.

MCCULLAGH, P. & NELDER, J.A. (1989). *Generalized Linear Models, 2nd edn.* John Wiley & Sons, New York, NY.

MORRIS, M.D. & EBEY, S.F. (1984). An interesting property of the sample mean under a first-order autoregressive model. *The American Statistician* **38**, 127–129.

NGUYEN, H., CRESSIE, N. & HOBBS, J. (2019). Sensitivity of Optimal Estimation satellite retrievals to misspecification of the prior mean and covariance, with application to OCO-2 retrievals. *Remote Sensing* **11**, 2770.

ROBINSON, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review* **15**, 351–357.

SPIEGELHALTER, D.J., ABRAMS, K.R. & MYLES, J.P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* John Wiley & Sons, Chichester, UK.

TALEB, N. (2007). *The Black Swan: The Impact of the Highly Improbable.* Random House, New York, NY.

TOBLER, W.R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography* **46**, 234–240.

WIKLE, C.K. & BERLINER, L.M. (2005). Combining information across spatial scales. *Technometrics* **47**, 80–91.

ZHANG, B., CRESSIE, N. & WUNCH, D. (2017). Statistical properties of atmospheric greenhouse gas measurements looking down from space and looking up from the ground. *Chemometrics and Intelligent Laboratory Systems* **162**, 214–222.

ZHANG, B., CRESSIE, N. & WUNCH, D. (2019). Inference for errors-in-variables models in the presence of systematic errors with an application to a remote sensing campaign. *Technometrics* **61**, 187–201.