ΠΕΔΟ METRON

The Newsletter of the Pedometrics Commission of the IUSS

Issue 32, December 2012 **Chair: A-Xing Zhu** **Vice Chair: Dick. J. Brus**
Coordinator: Murray Lark **Layout: Jing Liu**

From the Chair

Dear Fellow Pedometricians and Friends,

Tomorrow is December 21st, 2012 on which one of the three Ancient Mayans' calendar, the Long Count Calendar, will complete a major cycle. Some people have claimed that the date signifies the end of the world. There are many reasons that state this cannot be true. Pedometrics can contribute at least three.

The first is the wonderful stories in this issue of pedometrons, ranging from the spatialization of sins reported by Budiman and Alex, to guessing the variogram by Alex and his company, to Spiking to improve by Budiman and his contributors, to Murray's piece on Alan Turing the founding figure of digital computing. David's book review on "Field Sampling for Environmental Science and Management" by Webster and Lark keeps us in check with recent monographs in this field. The Pedomathemagica is always challenging and fun to read and to work on. The world cannot come to an end without these interesting, informative, fun pieces being read first.

The second is the exciting events to come in 2013 for pedometricians. In addition to many of the soil science society meetings in many countries, pedometricians have three major events to look forward to. The first is our own meeting, Pedometrics'2013 in Nairobi, Kenya, 26-31 August 2013. This is the first time for pedometrics to move into Africa. Many people are looking forward to this meeting. The second meeting is the IUSS Global Soil Carbon Conference to be held in Madison, Wisconsin, 3-6 June 2013. The Pedometrics Commission will host a session titled "Pedometrics: Understanding, mapping, measuring and monitoring soil carbon in space and time". I encourage pedometricians to take active part in this meeting. Besides, Madison in June is not only beautiful but also pleasant. Imagine yourself in shorts and slippers with soft (not just gentle) and cool wind caressing your skin while you are sitting on the terrace of Lake Mondota at the Memorial Union of the University of Wisconsin-Madison drinking the Spotted Cow and taking a bit at a Wisconsin brats. Come to join us. The third meeting is the Geomorphometry'2013 to be held in Nanjing, China, between October 16-20, 2013. Although it is not strictly a soils conference but digital terrain analysis is a key element in pedometrics and digital soil mapping and many of active players in pedometrics are also key players in geomorphometry. You might find some new friends there, too. Besides, Nanjing was ten times capital in history and is quite a place to visit. These and other meetings are just enough to stop the world from coming to an end.

The third reason is about the climate modelers whom we, the pedometricians, need to work on to stop them from bringing the world to the end. I have to continue this in the next issue because I am out of space here.

The Long Count Calendar completes a major cycle but the world will not stop here because it will continue at least for another major cycle over which Pedometrics will play a major role.

Merry Christmas and Happy New Years!

A-Xing Zhu, on December 20th, 2012 in Beijing, China

Inside this Issue

Did you miss this? *Sin spatialised*

By Budiman Minasny & Alex. McBratney
The University of Sydney

While pedometricians are busy with digital soil mapping, researchers from the Department of Geography at Kansas State University recently illustrated that mapping can be intriguing by analysing the spatial distribution of the seven deadly sins. Published in the Journal of Maps (http://www.journalofmaps.com/) in January 2012, authors Mitchel Stimers, Ryan Bergstrom, Thomas Vought, and Michael Dulin undertook the task of mapping the seven deadly sins at the county level within the Midwest region of the United States. The authors "*took a different approach, an approach that is sadly all too often forgotten in academia and scholarly pursuits in general, and that is one of intellectual curiosity." Perhaps pedometricians could start emulating this approach. The authors added "such research may not solve the world's ills, but if nothing else, it bespeaks of the possibilities which are inherent in the spatial sciences*".

Each of the sinful elements: Sloth, Greed, Envy, Gluttony, Wrath and Lust were calculated based on interesting interpretations of existing sociological and economic data. In another interesting interpretation, Pride, the "greatest" and "root" of all sins, was calculated as an aggregation of each of the sins and represents the total sinfulness of a given county or region.

The authors also performed a hot-spot analysis, showing areas of high, moderate, or low values through Gi* analysis which calculates the Z-score. The maps which can be downloaded from the authors' website (http://www.hazardgeographer.com/7_sins.html) showed that the area including and surrounding Kansas City, Missouri, has above-average GI* results (> 1.65 standard deviations) for six of the seven sins. Areas including and surrounding St Paul, Madison, Kansas City, Omaha, Lincoln, des Moines all show above-average GI* results in Greed. (These are all university towns!)

The authors also have mapped the seven deadly sins in the USA. Figure 1 shows that Wrath and Pride occurred most in southern USA. Except for Florida, the areas are dominated by the heavily weathered Ultisols. The results are interesting, and reasonably convincing. Can pedometricians do a better job on this since the first sin was done on soil? "*..., cursed is the ground for thy sake; in sorrow shalt thou eat of it all the days of thy life*". Finally, there is nothing in the report on uncertainty, and some of us think that is the biggest sin of all. But of course, let he who is without sin cast the first coarse fragment.

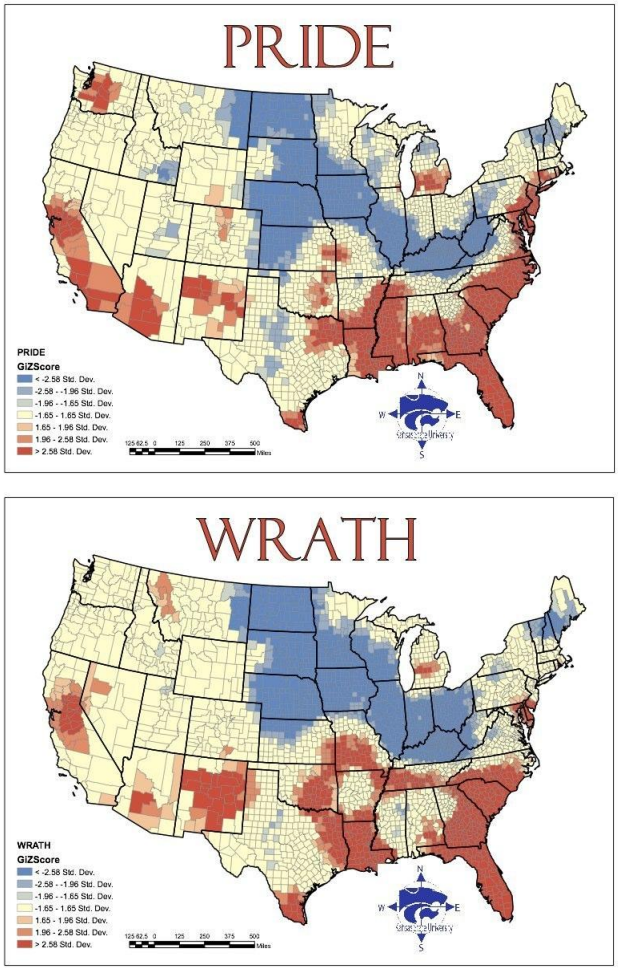


Figure 1. Hot spots for Pride and Wrath in the USA, with values corresponding to the Z score. Red areas indicate a standard deviation of 1.65 above the mean, blue areas indicate a standard deviation of 1.65 below the mean, and neutral coloured areas indicate values near the mean. Used with permission from Mitch Stimers, otherwise we would have created more sins.

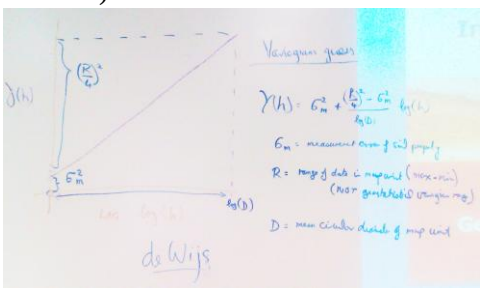
Reference

Stimers, M., Bergstrom, R., Vought, T. and Dulin, M. (2011) Published Map. In Stimers, M., Bergstrom, R., Vought, T. and Dulin, M. (2011) Capital Vice in the Midwest: The Spatial Distribution of the Seven Deadly Sins, *Journal of Maps*, v2011, 9-17. http://dx.doi.org/10.4113/jom.2011.1133

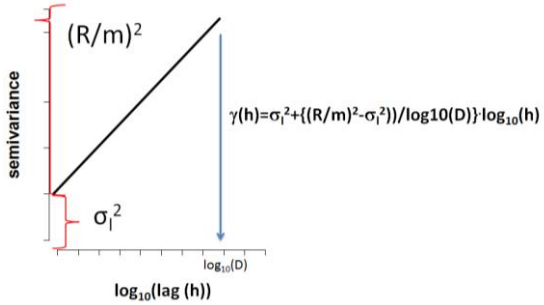
Have a go - guess the variogram

By Alex. McBratney, Budiman Minasny, Brendan Malone,
The University of Sydney

On August 28th or thereabouts at a meeting in Lincoln, Nebraska about estimating uncertainty for the GlobalSoilmap project Gerard Heuvelink reminded us that to achieve more sophisticated uncertainty estimates we would at some stage need to know the variogram of a soil property for a map unit or an area of interest. Often in reality we may have limited data to do this so we need some expert way of doing it. Alex. said that he had for some years had an informal way of doing this and he scribbled the following on the board (in his fairest hand, which we know is not all that flash).



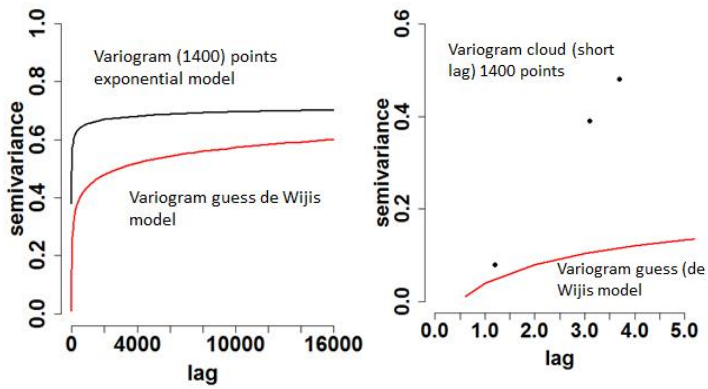
As this is largely illegible we reproduce it thus.



where σ_1^2 is the measurement error of the soil property. *R* is the range of data within mapping domain (max-min) and *D* is the mean circular diameter of the mapping domain.

Basically it is based on a belief that soil properties don't have real sills (just a product of the finiteness of areas of observation) and actually this makes life a lot easier – and we're all for an easier life. So for a soil property – we have a rough idea of the measurement error (method dependent of course) – we can tie that to say 1 m or some small lag unit of distance. The other information we need is the range of the data within a map unit or area of interest – let's say we know the maximum and minimum value, we then know the range (*R*) and a simple estimate of the variance is $(R/m)^2$, where *m* is the number of standard deviations that the data range roughly corresponds to, and this corresponds roughly to a lag equal to the equivalent circular diameter of the map unit or area of interest of area *A*, $D = 2\mathrm{sqrt}(A/\pi)$. If *m* is based on few observations, then $m = 4$. (If *m* is based on many observations, then $m = 6$ but if you have this many observations then you can probably estimate the variogram in the normal way.) We now draw a line from $\{1, \log_{10}(\sigma_1^2)\}$ to $\{D, \log_{10}((R/m)^2)\}$. This is the de Wijsian variogram model $\gamma(h) = \sigma_1^2+\{(R/m)^2- \sigma_1^2))/\log_{10}(D)\}\cdot\log_{10}(h)$. The use of the De Wijsian model for soil properties is suggested in Figure 1, McBratney, A.B. (1992) *Australian Journal of Soil Research* 30, 913-935. If the lag is now drawn on the linear scale of distance it appears to have a pseudo-sill at any scale. This is just a rough model to get started – don't assume we know nothing – in fact believe you know quite a bit.

Here is an example for topsoil pH (0-10cm) from the Hunter valley, NSW, Australia, the measurement error is about 0.02 pH units squared, the area is about 210 km², so *D* is about 16 km. The variance at lag h = D = $((7.9\text{-}4.9)/4)^2 = 0.6$. The range R comes from an observed maximum of 7.9 and a minimum of 4.8 based on 30 observations. This produces the following variogram (black line) – which we compare with a variogram based on ≈ 1400 data points (red line) (Figure left immediately below).



Not too bad. Guessing the short-range or nugget variance is tricky – the semivariance for the shortest lag from the 1400 observations is close to our modelled value (right graph immediately above) . You can consult Pringle & McBratney (1999, Precision Agriculture 1, 125-152) for a nugget variance for some soil properties, or use the measurement error or the measurement error times two to add in some spatial variance. Reported nugget variances from our community for a range of soil properties would be useful here. Try this out for your own area and properties – you might be surprised.

To spike or not to spike?

By Budiman Minasny and contributors

Introduction

Budiman Minasny (The University of Sydney)

A method called 'spiking' has been proposed in the recent soil science literature. It appears mainly in papers about calibration for soil infrared spectroscopy, but it has been suggested that it could be applied to other pedotransfer functions or predictive models more generally.

Spiking is concerned with the development of prediction functions that were calibrated on soil data from a regional or broader set for use in a local area. Prediction functions generated from data collected at a regional scale may not perform well in a local field. We have a large regional dataset relating, for example, near infrared spectra to soil carbon; and we want to use such data to produce a prediction function to use in a local area. Spiking involves the addition of a small number of samples from the local area to the larger regional dataset and the subsequent re-training or re-calibration of the model. It has been suggested that model developed based on this spiked or augmented data will provide a better prediction for the local area. It has also been proposed as a strategy to combine local and more general datasets for reducing the number of local samples and achieving more reliable predictions.

Several studies showed spiking improved the performance when compared to prediction with functions calibrated from the regional data set, mainly through a decrease in bias. Some people suggested that by adding a limited number of local samples the new model will cover the variation of soils not captured in the larger and more regional database.

A presenter at a conference once said that after spiking, the new model work miraculously for the local area. Do miracles happen in statistics? As pedometricians, we need a rigorous analysis of how and why before we can make a conclusion. We asked several pedometricians for their opinions.

Comments

David Brown (Washington State University)

We have utilized 'spiking' in the development of VisNIR calibrations and usually find improved predictions at a particular field with the inclusion of a few local samples in the calibration. As was noted in the introduction, improvements through spiking are usually (though not always!) related to a reduction in prediction bias. I don't think there is anything miraculous about the effectiveness of spiking. Chemometric modeling is an empirical approach that relies upon a similarity between calibration and prediction samples. For any kind of empirical, statistical modeling, interpolation is more reliable than extrapolation. For that reason, it is standard procedure in NIR food and grain analyses to continuously add samples to calibration sets as predictions are made for new batches, locations or seasons. The mineral composition of soils quite often varies from field to field (particularly at higher latitudes with greater soil mineral diversity), so we shouldn't be surprised that soil-VisNIR calibrations don't always transfer well from field to field within a particular region.

It would be more cost effective and convenient if we could minimize the number of "spiking" samples required. This could potentially be accomplished by screening new samples for spectral similarity to an existing soil-spectral database (as is routinely done in production NIR labs). The challenge for this approach is that (1) soil composition and spectra vary considerably more than most processed foods; (2) sometimes compositionally different soils can have similar spectra; and (3) small spectral differences can potentially result in major differences in the predicted response variable.

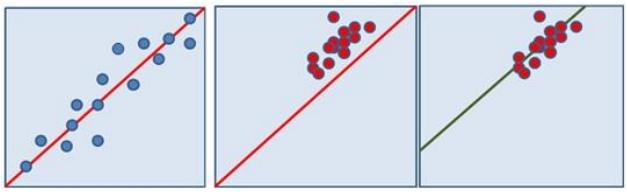
As a research community, we could also devote more effort toward the development of fundamental soil-spectral relationships. For example, Roger Clark and colleagues at the USGS Spectroscopy Lab (Denver, CO, USA) have derived a mineral identification tool based upon the measurement of mineral-specific band depths (adjusted for albedo). The challenge for soil scientists would be that we generally deal with much more complicated mixtures and less well defined materials than geologists.

Finally, it is worth noting that sometimes we are more interested in field-scale patterns than absolute values of targeted soil properties. In that case, a strong bias in predictions for a particular field may not be a problem as long as the spatial variability is well represented.

Bo Stenberg (Swedish University of Agricultural Sciences, SLU)

In our experience spiking has the potential to adjust a calibration from a regional library to better fit a local prediction area were the dependent variable, for

example organic carbon or clay content may vary enough, but the geological and management history is basically the same. What we have seen is that the bias is what mainly is corrected for and in cases where there is no bias spiking makes little or no difference (Wetterlind and Stenberg, 2010). The relationship, or ranking, between predicted values is markedly similar, spiked or not. The interpretation is that soils with compositional qualities, for example mineralogy and organic matter quality, not well represented in the regional data set will show larger residuals in the calibration. If a local dataset of this type is predicted, then the result could be biased. By spiking, this soil type is better represented, the calibration is adjusted and the residuals are reduced (see Figure 1).



(a) (b) (c)

Figure 1. A graphical representation on explanation the effect of spiking. The blue circles represent 'regional' data and the red circles represent 'local' data. (a) A regional model was fitted to the data represented by the red line, (b) the regional model's performance on local data where a more 'general' model may not be suitable, (c) by adding a few known local samples to the calibration(spiking) and recalibrating the model, the bias of the model is reduced.

Thus, there is nothing miraculous with spiking, but its effect can be substantial on validation statistics, especially if you compare with what is usually achieved with various spectral transformations or calibration algorithms.

Many issues regarding spiking need to be elucidated, for example questions about the method of selection and optimal size, and weight, of spiking sets. It can also be questioned if spiking is necessarily better than other alternatives. In our study (Wetterlind and Stenberg, 2010), we found that local calibration with as little as 25 calibration samples typically performed as good as or even better than a regional calibration spiked with the same 25 samples. There were, however, indications that spiking might perform almost equally with fewer samples.Another alternative could be to select only the most similar samples in the regional dataset for calibration, with or without spiking.

David Clifford (CSIRO Mathematical and Information Services)

"Spiking" a large database with a small amount of local data could well lead to an improvement in inference at the local level but that depends on how large, how small, and what kind of statistical inference is taking place.

Many non-spatial statistical models place equal weight on all observations by default, and in such a case, adding a small number of observations to a large database will have little noticeable effect. If the addition is immediately noticeable it could be for several reasons but none of them are good. The new data could be outliers relative to prior information (and vice versa). New data could also be points of high leverage indicating the covariate spaces covered by the database and the local area are quite different. Such incompatibility between information indicates that the use of the database for inference in the new area would be inappropriate.

For a process that varies spatially, and for inference procedures that are flexible enough to take advantage of such spatial relationships, the use of additional local data can be informative but good performance will be reliant on the overall spatial properties being maintained in this new area. This is true for full spatial models or even for models where a random effect is used specify a local mean or trend term. If that is all that is different about a new area then a small amount of data is enough to make improvements but it will depend on the process and data quality.

Finally, any model for the new area that is based on a small amount of additional local data will be evaluated using even fewer independent local validation points (or possibly via cross validation). Users should be wary of overfitting in such cases and be sceptical of miraculous model performance.

Gerard Heuvelink (Wageningen University)

I had never heard of spiking until Budi asked me about my opinion. As I understand it is about the choice between relying on a large dataset from a larger region and relying on a small dataset from the local study area for which predictions are to be made. Spiking offers some additional flexibility to mix both datasets and assign weights, but essentially it is about the trade-off between these two approaches. This is a difficult problem because we know that in the real world the stationarity assumption is never truly satisfied, which suggests that we should use models that are calibrated locally, but we also know that in

case of few observations model inference is poor. I don't think there is a universal solution to this problem, and although (cross-)validation may be useful because it will tell which of the two approaches works out best in a practical case, it is not always entirely valid and results cannot be generalized.

When comparing my personal choices with that of my peers I think I tend to go more often for the global approach. For instance, I prefer global kriging over kriging in a local neighbourhood and I am quite sceptic about methods such as moving-window kriging and geographically weighted regression. I guess it is because I prefer approaches that use an explicit statistical model as a starting point for which the assumptions are clearly defined, and do the inference and prediction in a way that automatically follows from the model that is assumed. The advantages of this are that ad hoc solutions are avoided and that the assumptions behind the predictions are unambiguously defined. If data are abundant I would obviously make use of the opportunity and relax the assumptions by assuming a more complex model, but I would keep the complexity within limits such that the resulting model remains estimable.

In conclusion, I don't think spiking is a method that I am eager to use but I do recognize its merits because it makes a less rigorous stationarity assumption.

Murray Lark (British Geological Survey)

As I understand it spiking entails the augmentation of a global calibration database with local observations prior to the estimation of parameters of predictive models. It is said that this has improved the quality of predictions in various circumstances.

One might expect that predictive models with parameters estimated from a calibration dataset collected in region A will generally be better than models with parameters estimated from a global data set for region A because in many cases these parameters may vary with climatic conditions, parent material etc. This isn't spiking, however. Spiking entails the use of the global dataset, but the effect is claimed to come from adding some local observations to this data set. It is said that this practice has been shown to work in a number of case studies.

Statistics is not an experimental science. That is to say, we do not establish what is sound statistical practice by empirical means. We establish what is sound statistical practice by theory. We may then use empirical studies to establish, for example, whether certain assumptions, such as stationarity in the variance or a normal distribution, are generally plausible in soil science; or to establish sampling requirements for particular tasks, but these start with a clear theoretical framework in the light of which information can be deduced from data. Studies which use spiking are different, they set out to establish empirically that the spiking technique is appropriate, and conclude that it may be useful on purely empirical grounds. I would argue that the findings of such studies are plausible but do not establish that spiking is an acceptable practice.

In some circumstances something like spiking might be sensible. Consider a situation where we have strong prior grounds to believe that a soil variable z is linked to another x by a linear relationship, and that this relationship holds over a very wide range of values of the two variables. We may hold one data set in which x falls in some interval $[x_{\min}, x_{\max}]$ and develop a PTF (or spectral calibration function) by linear regression. If we then wanted to use the PTF somewhere where for at least some observations $x < x_{\min}$ then we know that adding some observations in this range and refitting the PTF will improve the precision of predictions, that is because the expected mean-square error of a linear regression depends inversely on the dispersion of the values of x in the calibration set, as can be seen in the standard formula for the prediction error variance in any statistics textbook. An empirical study would be useful in the light of this theory, because it would provide the statistics that we need to estimate the gain in precision from using some new observations to refit the regression, and so to decide whether it is worthwhile. But we would know why there was a benefit, and know that there is a real benefit (it is not just a coincidence from random error in the formation of the data set that we have). Furthermore, we benefit in these circumstances by reducing the sampling error in our estimates of the regression parameters, not by reducing bias, so this is not the empirical effect commonly reported in spiking studies.

In other circumstances we can see why something like spiking might appear to improve predictions. Say that the relationship between z and x in one region is best described by the linear relationship $z = a_1 + b_1 x$, and in a second region $z = a_2 + b_2 x$. The two sets of coefficients may differ because of various factors that differ between the environments. Now if our original PTF is fitted to data from the first region, then it will do poorly at predicting for observations from the second, in so far as the parameters differ. If we added

some observations from the second region then, in general, we would expect the new PTF to be closer to the function for the second region and so to predict better. However, that is a rather arbitrary effect, and if we really understood what was going on we would chose rather just to fit the new PTF to data from the region in which we now want to predict. The somewhat better performance of the 'spiked' PTF would not justify the spiking approach.

If we approached the development of PTFs critically, then we would look at the distributions of the predictor variables and the predictands, and examine their functional relationships. That would tell us which of the two above situations pertained in any particular case, and we would act accordingly, either supplementing a common data set (first case) or, in the second case, producing a new PTF for the second region. One of the many good reasons to avoid data mining is the 'black-box' nature of the PTFs that are produced, which prevents their critical assessment. Having said that, the very least that we should do is to examine the data distributions (predictor and predictand) and avoid fitting PTFs for use in one data set with observations in a very different part of the predictor space.

In short, the spiking effect is not surprising, but to deduce, as many have, that spiking is the solution, is to fail to think critically about what is done in statistical prediction.

Summary

We have read the comments from a couple of proponents on spiking and three others which are not so keen on the method. Proponents of spiking mainly used it for soil infrared spectral calibration, with the aim to reduce the number of samples needed for calibration. They observed that there is generally a reduction in bias when the larger database was spiked. However Bosse also observed that spiking does not always work and that small local calibrations may do a better job.

The question remains how and why it works. Murray warned that this could be just an arbitrary effect, and we should first understand what's going on rather than try and see. While in spatial prediction, local data may be useful for additional improvement, the use of spiking is mainly for non-spatial calibration. As pointed by David Clifford, in most calibration model, adding few samples to a larger dataset shouldn't make a difference to the calibration model as the weights of the spiked samples are quite small (only 1-5%) compared to the regional dataset. Unfortunately mathematical spiking does not work like chemicals, where a small potent concentration of a chemical can affect a larger volume of solution. Then the question is when there is a noticeable difference in the calibration function, there is something going on. The spiked samples could be very different from the original dataset.

In simple linear model, spiking can be explained simply by looking at the covariate space of the predictors, the relationship between the spiked samples to the regional database. Most spiking technique is used in soil spectroscopic modelling where the model is a high dimensional multivariate model. While Partial Least Squares model is still linear, it is made up of few hundreds of predictors, and difficult to see if the model overfits the data. Nevertheless we can still examine the covariate space of the predictors, by simply plotting the first two principal components. As an example we see from Figure 2, a plot of the first 2 principal components of soil near infrared spectra (Figure courtesy of Fan Deng, Aarhus University). The black dots represent the Danish national database which contained about 2800 samples, which were scanned under dried and ground conditions. The blue triangles are the principal components from the spectra obtained from a field called Vindum in Denmark, the 36 samples were scanned in field conditions. Spiking would involve mixing the Vindum samples into the national data. We can see that the two datasets occupy different parts of the principal component space. And the Vindum spectra exhibit a large variation. Does it make any sense to combine these 2 incompatible datasets hoping that it will make a better model? In this case, we should remove the effect of moisture from the spectra. There is a chemometric procedure that does that.

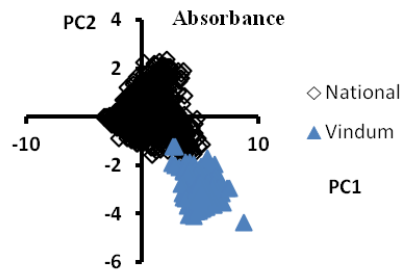


Figure 2. Plots of the first two principal components of the near infrared absorbance spectra. The black diamond represents the Dansh national dataset (2500 samples, scanned in ground and dried condition). The blue triangles are spectra from a field called Vidum (in Denmark) that were scanned in the field condition. (Figure courtesy of Fan Deng, Aarhus University).

To spike or not to spike?

In summary, while spiking works in some cases, we should be critical and establish good science based on theoretical grounds. We shouldn't based it empirically by trial and error to see if it works. How will we know it will work if we make a prediction on a new field where there is no validation? We should at least see where the predictors lay in the covariate space and then make a judgement.

References

Wetterlind, J. and B. Stenberg (2010). Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science* 61(6): 823-843.

Alan Turing, Statistics and Pedometrics

By Murray Lark
British Geological Survey

Pedometron cannot let the year 2012 pass without noting the centenary of the birth of Alan M. Turing, the British mathematician whose contributions to the development of digital computation make him a founding figure of the modern age. Turing was a mathematician, and his PhD and associated work was on mathematical logic. His best-known contribution was on the Entscheidungsproblem attributed to Hilbert: is there a definite method in arithmetic that can be applied to any proposition to decide whether it is true? Turing tackled this by developing the concept of computable numbers. A computable number can be obtained by implementing what we would now call a program on a Turing machine, a hypothetical device that processes numbers held on a tape that serves as its memory according to its particular configuration (which may be changed depending on the numbers it finds). What Turing showed is that there must be uncomputable numbers, and this settled Hilbert's problem.

So far, so abstract. The point was that Turing, despite his interest in these theoretical problems, was also deeply interested in real machines. When the Second World War broke out he became involved in the British Foreign Office's Cryptography school, based at Bletchley Park (not very far from Rothamsted, to the north of London). It soon became apparent that some of the properties of Turing machines could be realized in very real computing machinery and applied to otherwise intractable cryptographic problems. The particular challenge was to decrypt intercepted messaged that had been encrypted on the Enigma machine (Hodges, 2012).

Well, pedometricians use computers directly descended from those that Turing and his colleagues developed, but have we any other reasons to be interested in Turing? In fact there are two particular areas of statistics to which he contributed. As an undergraduate he came up with a proof of the central limit theorem by which the aggregation of many small random effects is a normally distributed random variable. He was just scooped by another worker, but wrote his work up for a thesis which secured him a fellowship of Kings College, Cambridge. The second area in which Turing contributed to understanding of statistical inference was his work on quantifying the weight of evidence.

Weight of Evidence

Imagine that you have a choice between two alternative, mutually exclusive values for the underlying state of a system, p and ¬p. We can assign prior probabilities to these states, $\pi(p)$ and $\pi(\neg p)$ respectively. Clearly $\pi(p) = 1 - \pi(\neg p)$. From Bayes's rule we can compute the posterior probability for state p given some data, D: $\pi(p|D)$.

$$\pi(p|D) = \pi(p)\,\pi(D|p)/\,\pi(D).$$

Where $\pi(D|p)$ is the probability of the data given the state (i.e. the likelihood of the state) and $\pi(D)$ is the probability of the data. Turing did not work in terms of probabilities but rather the odds. The odds of the system being in state p are given by the ratio $\pi(p)$: $\pi(\neg p)$. Now, from Bayes's rule we can see that the

posterior odds ratio, $\pi(p|D)$: $\pi(\neg p|D)$ is given by $\{\pi(p)/\pi(\neg p)\}\{\pi(D|p)/\pi(D|\neg p)\}$. That is to say, the posterior odds is equal to the prior odds times the ratio of the two likelihoods. This latter term is now called the Bayes factor, *K*. The Bayes factor is a measure of the strength of evidence for state p. It is independent of the priors (which may vary from person to person), but indicates how the priors should be rationally modified by the evidence provided by D. The idea that this ratio is a measure of weight of evidence was proposed by Turing and his colleague at Bletchley Park, I.J. Good, but not explicitly in Bayesian terms (Good, 1979).

Bayes factors have been quoted at least once in the Pedometrics literature (Orton et al., 2011). Jeffreys (1961) proposed a scale for the interpretation of the Bayes factor. If $3<K\leq 10$ then the evidence for state p is said to be substantial. If $K \geq 100$ then the evidence is said to be decisive.

The deciban

Turing wanted to assess the evidence that intercepted messages provide for the underlying state of the Enigma machine. Once this information was obtained messages could be decoded, until the state of the machine was changed again. One semi-mechanized procedure to extract information from intercepts entailed the comparison of two messages to identify phase shifts at which letters matched more frequently than is expected for random sequences. This process was named Banburismus at Bletchley Park. This is because Banburismus used punched cards to detect interesting shifts, and these were printed in the nearby town of Banbury.

Observed shifts in Banburismus might offer evidence for two contrasting states of the Enigma machine. The evidence can be weighed by looking at the [Bayes] factor (i.e. the likelihood ratio) for the different states. If another pair of messages are then compared additional evidence may be gathered. The updated odds ratio for a particular state is obtained by multiplying the posterior odds after the previous evidence by the likelihood ratio from the new observations. However, this gets messy, what we are really interested in is not the odds ratio (which depends on the priors) but the accumulated evidence. A natural way to look at the accumulating evidence for the different states is to take logarithms. The log of the posterior odds ratio, after several rounds of observations, is equal to the log of the prior ratio plus the sum of the log likelihood ratios for each piece of evidence. This sum of logs of likelihood ratios is Turing's proposed weight of the evidence. He used logs to base 10, and the unit of evidence on this scale is the Ban (after Banburismus) — in practice one works with deciBans (10-1 Bans) so substantial evidence, on Jeffreys's scale is 5–10 dB. As is often the case, this innovation was not entirely new; the philosopher Charles Peirce had proposed that weight of evidence be measured by just such a log-ratio in 1878 (Good, 1979). It has been suggested that 'a deciban, or half deciban, is about the smallest change in weight of evidence that is directly perceptible to human intuition' (Good, 1979). Good suggests that it might be usefully applied by 'doctors, lawyers or other citizens'. Maybe also by pedometricians, Orton et al. (2011) provide an example, evaluating evidence for contrasting assumptions about the kinetics of denitrification in soil in different sections of a transect in terms of categories of weight-of-evidence about 5 dB wide.

References.

Good, I.J. 1979. Studies in the history of probability and statistics. XXXVII. A.M. Turing's statistical work in World War II. *Biometrika* 66, 393–396.

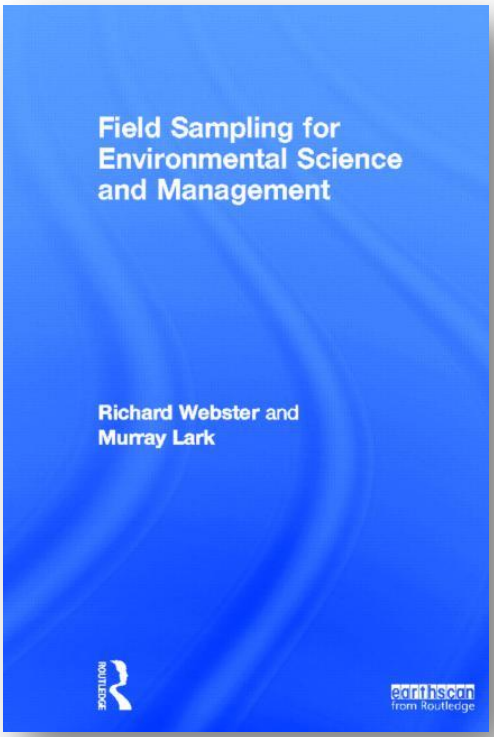
Hodges, A. 2012. *Alan Turing, the enigma*. Vintage, London.

Jeffreys, H. 1961. *Theory of Probability*, 3rd edition Oxford University Press, Oxford.

Orton, T.G., Goulding, K.W.T., Lark, R.M. 2011. Process model comparison and integration in a geostatistical case study incorporating the opinions of process experts. *European Journal of Soil Science* 62, 359–370.

Book Review on "Field Sampling for Environmental Science and Management"

By D G Rossiter
University of Twente



Webster, R., & Lark, R.M. (2013). *Field Sampling for Environmental Science and Management*. Abingdon, Oxon (UK): Routledge.

http://www.routledge.com/books/details/9781849713672/ Also available as e-book and Kindle book

Book reviews become less relevant in the digital age, since for this book and many others readers can form their own opinions with the very nice "look inside the book" from the Amazon on-line bookseller (http://www.amazon.com/Field-Sampling-Environmental-Science-Management/dp/1849713685), which shows the table of contents, index, and selected pages. Still, that does not stop me from giving my opinion in this review.

The authors position this book as a bridge between the statistics taught in typical environmental science courses and what the graduate would then need to know about sampling to make good decisions on sampling design. It is consciously less comprehensive, less expensive, and, the authors hope, more accessible than the new standard reference on the topic, de Gruijter et al. (2006). The authors are well-known, to say the least, in pedometrics, and the senior author has produced excellent didactic material for 35 years, most notably (to this reviewer) the 1990 "Statistical methods in soil and land resource survey", which unfortunately has been allowed to go out of print. Most readers of this newsletter would be interested in the present offering as a textbook for their students or a self-study book for new research associates with weak backgrounds in statistics. Each chapter begins with boxed "key points", and there are numerous case studies with real datasets to illustrate how the theory works out in practice. From the didactic point of view, especially for self-studiers, it is a real pity that these datasets and the analytical procedures are not provided as supplementary material (as for example Tom Hengl's work and in the UseR! series. e.g., Bivand et al.). Finally, I miss a summary aide-memoire, listing the key decisions to be made during sampling design and referring back to the main text, as Webster has provided in previous texts.

What do I make of this book? I begin with what is, to me, a misleading title "Field sampling...", reinforced by a cover photo of a power core sampler, implying that the book will be mainly about how to carry out sampling in the field. There is absolutely nothing about that here. One added word would have cleared up the confusion: "Planning...", and the cover should have shown a sampling design. Second, even the rest of the title is somewhat misleading, in that the book covers several purely statistical issues (t-distribution, simple linear regression, ANOVA) at an elementary level, presumably to lay the groundwork for discussing sampling designs that optimize inferences from these. However, these sections are not at useful level of detail comparable to the 1990 book; readers who have not studied these in an introductory statistics course will not find the explanations here sufficient, and those who have will skip over them. The introduction to geostatistics is more justified, since that is rarely covered in a first environmental statistics course. I would have preferred a book purely on sampling design, and a reprint of the 1990 book plus references to the 2008 geostatistics book for the theory.

There is certainly useful information here. I especially liked Chapter 4 "Efficiency, economy and logistics", where the authors build up an approach to extracting

maximum information at minimum cost; however, this is on the basis of uncertainty tolerance, not an economic loss function. Especially interesting to me was the section of ranked-set sampling, where the number of determinations (not samples) is fixed, usually because of their high cost. Chapter 8 is yet another presentation of a Webster favourite, nested sampling, updated now with REML estimates of the variance components. Here the mathematical level takes a huge leap over the extremely simple approach of the previous Chapter 7 on sampling for regression, where there is not a matrix let alone a likelihood in sight. Surprisingly, structural relations are not mentioned in that chapter, despite Webster's two excellent articles (1989, 1997) on the subject. Chapter 5 on prediction from spatial classifications brings the expert soil surveyor back into the sampling plan -- if we can properly stratify the landscape into classes with reduced within-class variance and place unbiased purposive samples in them! The problem is, we have no way of objectively knowing how successfully we achieved this objective.

In summary, experienced pedometricians should leaf through the book, looking for interesting details that may have escaped them; however their time is better spent with de Gruijter and colleagues. They should also evaluate whether the book is useful for its stated didactic purpose, in their context. I still hope for the 1990 text to be reissued; it gives a much sounder theoretical basis, with abundant practical advice, on which to base further study and practice. I would have liked to see the current book implemented as an extension of that work.

References:

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer. http://www.asdar-book.org/

de Gruijter, J., Brus, D. J., Bierkens, M. F. P., & Knotters, M. (2006). Sampling for Natural Resource Monitoring. Springer. IBSN 978-3-540-22486-0. Also as e-book.

Hengl, T. (2009). A Practical Guide to Geostatistical Mapping. Amsterdam. http://spatial-analyst.net/book/

Webster, R. (1977). Quantitative and numerical methods in soil classification and survey. Oxford: Clarendon Press. ISBN 978-0198545125

Webster, R. (1989). Is regression what you really want? Soil Use & Management, 5(2), 47–53.

Webster, R., & Oliver, M. A. (1990). Statistical methods in soil and land resource survey. Oxford: Oxford University Press. ISBN 978-0198233169

Webster, R. (1997). Regression and functional relations. European Journal of Soil Science, 48(3), 557–566.

Webster, R. (2001). Statistics to support soil research and their presentation. European Journal of Soil Science, 52(2), 331–340.

Webster, R., & Oliver, M. A. (2001). Geostatistics for environmental scientists. Chichester: Wiley & Sons. ISBN 0-471-96553-7

Webster, Richard, & Oliver, M. A. (2008). Geostatistics for environmental scientists. John Wiley & Sons Ltd. ISBN 978-0-470-51726-0. Also as e-book.

PEDOMATHEMAGICA

Problem 1 (easy)

Philippe, a French winemaker, has two barrels of wine that each contain 20 litres of wine. One barrel has red wine, the other white wine. Philippe decides to make rosé wine by mixing the wine from the two barrels (even though French law does not allow this!). He takes one litre of wine from the 'red' barrel and adds it to the 'white' barrel, mixes this very well, takes one litre of wine from the mixture and pours it into the barrel with red wine. Intrigued by this ingenuous way of making rosé wine, he wonders: is there more red wine in the white barrel or more white wine in the red barrel? Once you solved this problem, can you also tell how often the mixing should be done in order to ensure that each barrel has between 49 and 51 per cent of red and white wine?

Problem 2 (medium)

After solving Problem 1 Philippe decides to have a few drinks with his friend Marc. He has had quite enough wine, so they decide to drink beer instead. Both Philippe and Marc like drinking beer and like solving puzzles so they decide to play a game. The 21 glasses of beer that they wish to drink are placed on six horizontal lines as shown in the figure below. They agree to take turns and that in each turn they can take and drink as many glasses from any one of the rows as they wish (but from just one row and at least one glass). The one who takes the last glass must pay the bill. Philippe is the first to start. Can you advise him a strategy that ensures that he will have free drinking?

(From Gerard Heuvelink)

PEDOMATHEMAGICA

Problem 3:

Alf and Bert the soil surveyors go to the pub for a Christmas celebration with their colleagues Charlotte, Detleef, Englebert, Farai, Guinevere and Hai. They are disturbed by a visit from some carol-singers who are performing the traditional English carol:

> *On the first day of Christmas my true love sent to me; a partridge in a pear tree.*
>
> *On the second day of Christmas my true love sent to me; two turtle doves and a partridge in a pear tree*
>
> *On the third day of Christmas my true love sent to me; three french hens, two turtle doves and a partridge in a pear tree....*

At this point the soil surveyors drive the singers away with a well-aimed volley of bar snacks. Once peace is restored Alf says "That song is actually quite interesting. The singer receives one present on the first day of Christmas, three on the second day, six on the third and so on. I'll buy a pint for anyone who can tell me how many presents she would receive on the 365^{th} day of Christmas".

Guinevere, who is a dab hand with R, generates the correct answer with a simple algorithm on her laptop which generates the sequence of answers for n =1, 2, 3, ..., 365 days. Alf is not impressed by Guinevere's numerical cheating, which is certainly not in the spirit of Pedomathemagica. So what is the general solution to the number of presents received on the n^{th} day? (The answer that there are only 12 days of Christmas is not in the spirit of Pedomathemagica either).

Alf is not used to buying pints of beer to pay back bets (particularly for a simple warm-up Pedomathemagica like that). So he comes up with another question. "How many of us are here?" he says, "Ah, eight. O.K. Guinevere. You go to Flossie the bar-maid and ask her to choose any two odd integers, the only condition being that they are different. Square both numbers and subtract the smaller result from the larger. Order that many pints. If those pints can be shared equally among all eight of us without remainder, then I win the bet and you pay for the round. But if the remainder is anything from 1 to 7 pints, then you win the bet and I shall pay." Guinevere thinks: "In one case out of 8 possibilities I lose, in 7 cases I win. That's a good bet!" and she agrees. What is the probability that Alf wins the bet? Why might Guinevere have made a better decision if she had answered the first question properly?

(From Murray Lark)

PEDOMATHEMAGICA

Answers to the previous Pedomathemagica

Problem 1 From Murray Lark

Bert will be buying all the beer and digging all the pits. It may seem obvious that $0.999\dot{9} \neq 1$ because "it never quite gets to 1". However, if it is true of two real numbers x and z that $x \neq z$, then I must be able to order them, so that, for example, $x < z$ and furthermore I must be able to specify a third number, y, such that $x < y < z$, but what real number could Bert quote that falls between $0.999\dot{9}$ and 1.0? In fact, $0.999\dot{9} = 1$ which we can show, let $0.999\dot{9} = x$, then

$$10x = 9.999\dot{9}$$
$$= 9 + x$$
$$\text{so } 9x = 9 \text{ and } x = 1$$

A more rigorous proof that our system of writing numbers does not have unique representations of all numbers (1 and $0.999\dot{9}$, denote the same number, so do 2 and $1.999\dot{9}$ etc) was first given by Euler. It is based on recognizing that the terms of $0.999\dot{9} = 9/10 + 9/10^2 + 9/10^3 + \ldots$ are in a geometric progression, and finding the sum; that is an exercise for the reader.

Problem 2 From Gerard Heuvelink

We use two tables to solve this problem:

	chairman	secretary	treasurer
Dr. Phil			
Dr. Jack			
Dr. Sheila			

	Chicago	New York	halfway
Phil			
Jack			
Sheila			

Premise 1 tells us that Sheila lives in Chicago, so we can fill in the corresponding row of the table. Premise 4 tells us that someone earns exactly three times as much as the secretary of the PM commission, and that this person is his neighbour. This cannot be Sheila, because she lives in Chicago. It also cannot be Jack, because $40,000 cannot be divided by 3 exactly. So it must be Phil who lives half-way between Chicago and New York, and hence Jack lives in New York. One table is filled completely. We can now deduce from premise 5 that Dr. Jack is the secretary of the PM commission. Premise 6 tells us that Dr. Phil cannot be the chairman of the PM commission and must therefore be the treasurer. Dr. Sheila is the chairman.

Problem 3 From Gerard Heuvelink

Not long after Pedometron 31 came out I got an email from Brendan Malone from the University of Sydney, writing:

Hi Gerard,

The Thiessen polygon problem has really captivated me and presents an interesting challenge. It is doable but certainly it is taking much time for my algorithm to converge to the optimum. In the meantime here is a sample configuration which i know is very close to the optimum. Is it possible for you to see how close this is to the true configuration? I will send you the algorithm shortly which in essence is based on a simple annealing schedule (well that's what i think it is anyway). My algorithm is still trying to converge to the optimum and has been going for many hours. I may need to make some minor adjustments to fine tune the efficiency in which the optimum is reached.

Now i better get back to my real work...

Cheers,

Brendan

PEDOMATHEMAGICA

Answers to the previous Pedomathemagica

Problem 3 (continue)

Brendan had looked for a numerical, iterative solution to the problem: start with an initial solution of the 25 points distributed in space (first guess, anything is possible); calculate the value of a cost criterion (such as the number of grid cells for which the nearest neighbour map of the current solution does not agree with the given map), perturb the solution a little bit by moving one of the points in a random direction; calculate the cost criterion again; accept the new solution if it reduces the cost criterion and accept it with some probability if it increases it (to escape from local minima); repeat this procedure thousands of times and hope that it converges to the optimum.

I checked Brendan's solution and had to disappoint him: the positions of the 25 points of his solution were on average off their true location by 30 times the grid mesh, while the condition was that each point should be no more than two times the grid mesh away from their true location. I also told Brendan that he was in competition with someone else, who had mailed me that he was close to a solution but did not send it to me. So Brendan intensified his work (sleepless nights) and was able to improve his algorithm and reduce the positional error to an average of 10 times the grid mesh. Still too large and indeed in the meantime his competitor, Jasper Vrugt from the University of California, Irvine, had solved the problem. Jasper's solution had a positional accuracy of 0.9 times the grid mesh, which was well within the prescribed limit. Congratulations to Jasper, the bottle of Champagne is yours!

As many of you will know Jasper is a champion in numerical optimisation and has developed and published about numerous optimisation algorithms over the years, with exotic names such as SCEM-UA, DREAM and SODA. For this problem he used the DREAM algorithm, which is the acronym of DiffeRential Evolution Adaptive Metropolis, for this problem. For details see Vrugt, J.A., C.J.F. ter Braak, C.G.H. Diks, D. Higdon, B.A. Robinson and J.M. Hyman (2009), Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, Int. J. Nonlin. Sci. Num. 10, 271–288. It is much more intelligent than straightforward simulated annealing but we will have to ask Jasper to explain how it works exactly. In any case the figure below shows that it works very well.

So now we have the proof that we can reconstruct the original data from a map that was created by nearest neighbour interpolation. Be careful with publishing such maps because in effect you give away your data!

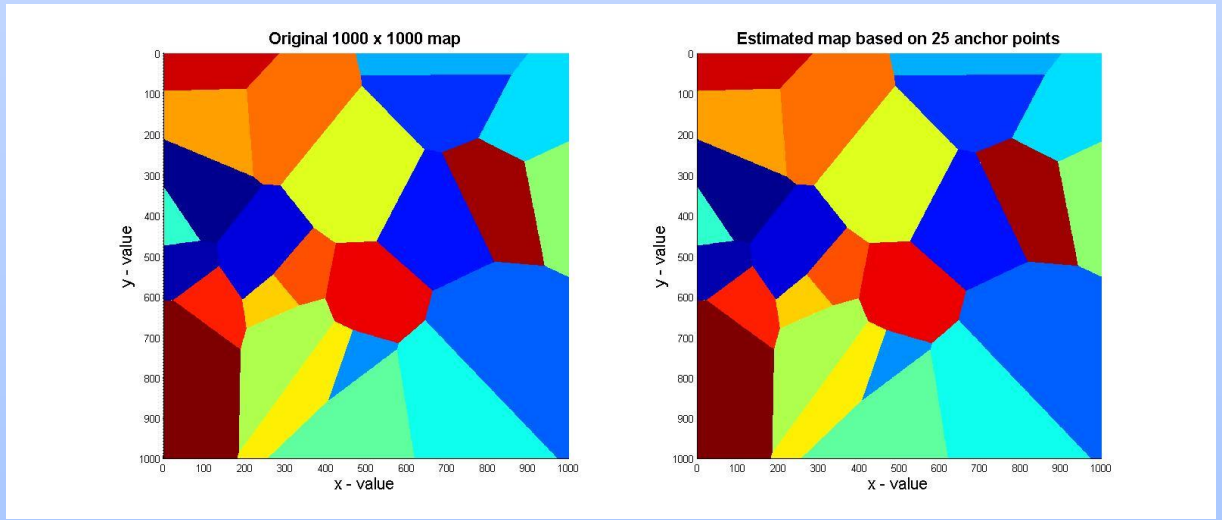


Figure: Original map (left) and Jasper's solution (right). He got the 25 point locations correct with minimal positional error.