REPORT FROM THE TRENCHES:

Preparing developing-country students for pedometrics

D G Rossiter

International Institute for Geo-Information Science & Earth Observation (ITC)
Enschede (NL)
http://www.itc.nl/personal/rossiter/

Pedometron and journals such as Geoderma are full of exciting and sophisticated developments in the application of math and statistics to soil science. In most less-developed countries, however, these are hardly known or mechanically and often inappropriately applied. There is a serious disconnect between these two worlds, which ITC is charged with bridging as part of its mission: "capacity building and institutional development of professional and academic organizations and individuals ... in countries that are economically and/or technologically less developed." The "capacity" in this context is the ability to understand and apply pedometric techniques for a deeper understanding of the soil resource and to make better decisions.

All ITC students are post-graduate and most are supposed to have some working experience in their professional field. They come to the Netherlands to upgrade their skills and apply them in an MSc thesis. In fields such as earth sciences they are assumed to have an appropriate university degree, which should include the relevant domain background (e.g. geology) and also relevant methods (e.g. statistics and university-level maths). Unfortunately almost all our students are deficient in one or both of these areas. Yet, we want to educate them and thus contribute to development. Good examples are an agricultural statistician from Malawi and an urban planner from China, neither who has taken a soils or even earth science course at university and with no soils field experience, who have been assigned by their respective ministries to learn about soils to apply in their jobs. Other students have some soils background and work experience but almost no statistics, let alone calculus or linear algebra; this is typical of agriculture college graduates in many developing countries.

ITC has not had a separate soil survey course for several years; in common with many universities soils are now included somewhat vaguely in earth sciences, natural resources, water resources, and even urban planning courses. All of these require sound statistical thinking, especially for MSc thesis research. In our modular system all MSc students are exposed to statistical thinking in a Research Skills module, and are offered optional advanced topics in data analysis strategy, geostatistics, and quantitative modelling. Domain knowledge such as soil science is insinuated when possible, mostly via directed readings assigned by the student's tutor and thesis coach.

(ITC also offers distance education courses, for example my "Geostatistics and Open-Source Statistical Computing", six weeks half-time; here I only deal with the MSc course.)

What do I do with these students, in the limited time, and given the impossibility of a semester course or sequence?

Above all, I want them to learn how to learn:

(1) They should be able to read and understand statistics textbooks in order to apply the right techniques for each situation, and meet the assumptions of each technique.

I expose them to a variety of texts available in our library, and show how to pick one at the level appropriate for them. For most earth science students the text of Davis is at the perfect level, and contains a wide variety of relevant topics:

Davis, J.C., 2002. Statistics and data analysis in geology. John Wiley & Sons, New York, xvi, 638 pp.

For soils students that are bit more sophisticated I recommend:

Webster, R. and Oliver, M.A., 2008. Geostatistics for environmental scientists. John Wiley & Sons Ltd., 332 pp.

although I think the earlier text is more useful for beginners; too bad it's out of print and the publisher won't let us photo-copy it:

Webster, R. and Oliver, M.A., 1990. Statistical methods in soil and land resource survey. Oxford University Press, Oxford.

The book of Goovaerts is an excellent and comprehensive reference but too detailed for most beginners:

Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Applied Geostatistics. Oxford University Press, New York; Oxford, 483 pp.

A problem with texts for our clients is their price. I have tried without success to negotiate with the publishers of Webster and Goovaerts to either buy books at substantial discount or photocopy them and send the royalties to the publisher; one publisher offered a 10% discount and the other never answered. Publishing on-line or as e-books may be a solution: the UseR! series is somewhat more reasonable, e.g. $60 for:

Bivand, R.S., Pebesma, E.J. and Gómez-Rubio, V., 2008. Applied Spatial Data Analysis with R. UseR! Springer, 378 pp.

(2) Students should be able to understand journal articles and repeat the methods on other datasets.

Often the students must review concepts presented in the paper that they do not know. Here the reference list is quite important, as well as a clear expository style.

A good example of an accessible paper (among many I could have chosen) is:

Minasny, B. and McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma, 142(3-4): 285-293.

Their section on Theory is a clear exposition of the choices they made, and why, with references appropriate for a student without the necessary background. For example: "Training in supervised classification involves minimising some error measure (Hastie et al., 2001)", the cited reference is a good text:

Hastie, T., Tibshirani, R. and Friedman, J.H., 2001. The elements of statistical learning : data mining, inference, and prediction. Springer series in statistics. Springer, New York, xvi, 533 p. pp.

This is followed by a clear derivation.

Many developing-country workers do not have good library access, either physical or internet. Thus references should be as accessible as possible. Papers from conferences or obscure journals (unlikely to be available) should be avoided if possible.

(3) It is more important that students understand statistical thinking, rather than specific methods.

All statistical models have assumptions: what are these? how can you tell if they're met? what are the consequences of violating them?

For example, kriging interpolation is applicable in the presence of stationary spatial dependence which can be modelled, but if the geographic phenomenon is due to a regional trend, it is certainly not appropriate. So I spend considerable effort in comparing approaches and when each may be applicable.

(4) Some fundamental methods must be understood in some detail; the most important is linear modelling (single and multiple predictors) in feature space (also, trend surfaces in geographic space, although these are less useful).

For geostatistics, the fundamental methods remains trend identification and removal, variogram analysis and ordinary kriging.

Computer programs

For our client group I insist on free computer programs. Fortunately one of the best is not only free but open-source: the R environment for statistical computing (http://www.r-project.org/). I prepare all my exercises with R, Sweave and LaTeX so that the executable code is provided along with verified output. An outstanding feature of R is the wide variety of contributed packages, so the student soon sees "there's more than one way to do it". Also, methods typically have many options, all of which are applicable in some situations. The student learns that "press the button" or "accept the defaults in a dialog box" is not acceptable practice. Life is complicated, accept it!

R is also fairly easy to program, and is based on a modern programming language. I have prepared some technical notes (available via my ITC home page) using R, e.g. implementing Webster's split-moving-window approach:

Webster, R., 1973. Automatic soil-boundary location from transect data. Mathematical Geology, 5: 27-37.

The more ambitious students are able to write simple programs or modify existing ones.

I avoid Excel (or open-source equivalents) for anything beyond initial data entry; far better to get the data into R and develop analysis scripts which allow reproducible analysis and professional graphics.

Commercial programs such as SPSS and ArcGIS Spatial Analyst have three strikes against them for the group I am trying to teach: cost, push-the-button ease of use, and poor programmability. Spatial Analyst is very poorly documented; despite repeated attempts I have not been able to discover how the empirical variogram display is computed nor how a variogram is fit.

Papers

Finally, here is a list of some of my favourite journal articles for teaching. I have an extensive list of specialised papers from my favourite pedometric authors (e.g. Lark, Minasny, Viscarra Rossel, Brus) which I recommend to students as they enter their thesis phase; these are more general and used in teaching.

(1) Statistical thinking and elementary methods

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

This one is simple but so many students benefit from just such an approach. This is supplemented by the aide-memoire from the Webster & Oliver text listed above.

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

Far too many students jump into regression when it's structural relations they really want. I find Webster's expository style a good model for the students.

(2) Geostatistics

Oliver, M.A. and Webster, R., 1991. How geostatistics can help you. Soil Use & Management, 7(4): 206-217.

This is the most gentle introduction to "why should I learn this complicated stuff?".

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma, 103(1-2): 3--26.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma, 89(1-2): 1-45.

Both of these are comprehensive comparisons of approaches.

Webster, R., Welham, S.J., Potts, J.M. and Oliver, M.A., 2006. Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. Computers & Geosciences, 32(9): 1320-1333.

I have a soft spot for this one, since I grew up in Youden and Mehlich territory (upstate New York) and have visited their study area.

(3) Case studies

Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology, 228(1-2): 113-129.

Dubois, G., Malczewski, J. and Cort, M.D., 2003. Mapping radioactivity in the environment - Spatial Interpolation Comparison 97. EUR 20667 EN, Office for Official Publications of the European Communities, Luxembourg.

This serves as a model of an intelligent approach to solve a problem with a variety of techniques, pointing out the (dis)advantages of each. Papers have been collected in an EU publication free for download.

SICC '97 http://www.ai-geostats.org/index.php?id=45

(4) Digital soil mapping

Anyone getting into DSM is given this one, of course:

McBratney, A.B., Mendonça Santos, M.L. and Minasny, B., 2003. On digital soil mapping. Geoderma, 117(1-2): 3-52.

Conclusion

Pedometricians, keep on inventing the latest sophisticated methods! Keep on publishing excellent papers and writing reviews and texts. However, spare a thought for those who are far below the level needed to appreciate your cutting-edge work, and provide a stepped approach to bring them into the community.