EnviroInfo 2007 (Warschau)
Environmental Informatics and Systems Research
 ISBN: 978-3-8322-6397-3

Directional Adjustment of Average Rankings of Two Posets: an Application to Herb Layer and Moss Layer Pollution in Baden-Württemberg

Michael Rademaker[1], Bernard De Baets[1], Hans De Meyer[2], Rainer Brüggemann[3]

Abstract

In the paper two procedures for Directional adjustment of average rankings of two posets: an application to herb layer and moss layer pollution in Baden-Württemberg are disscussed.

1. Introduction

We consider pollution data coming from 59 regions in Baden-Württemberg (BW) (see Brüggemann et al., 1998, Brüggemann et al. 1999, Pudenz et al., 1998). Each of these regions has been examined for pollution on three different layers: moss, herb and tree layer. In the herb layer, for example, the Zinc (Zn), Sulfur (S), Cadmium (Cd) and Lead (Pb) contents have been measured at a representative site in each of the 59 regions. On the basis of these measurements, a partial order can be built w.r.t. each of the three layers (Hasse diagram technique, see for example Brüggemann et al., 2001). Furthermore, the 59 regions can be grouped according to common properties, such as soil type.

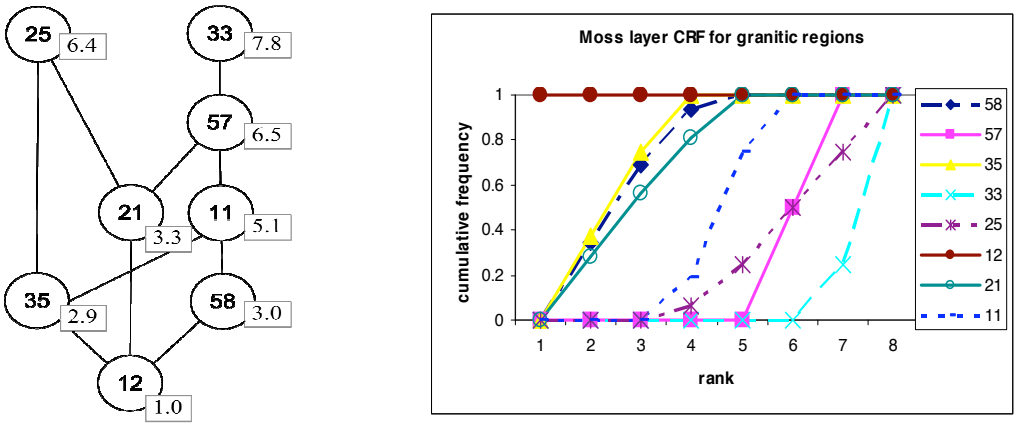

Figure 1: Moss layer HD showing average ranks and CRF for granitic regions

[1] Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Gent, Belgium, e-mail: {Michael.Rademaker Bernard.DeBaets}@ugent.be
[2] Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium, e-mail: Hans.DeMeyer@ugent.be
[3] Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Department Ecohydrology, Müggelseedamm 310, 12587 Berlin, Germany, e-mail: brg_home@web.de

To give an example, in Figure 1 we show the Hasse diagram (HD) of the Cd and Pb pollution data of the moss layer for 8 regions in BW with a granitic soil composition.
The most heavily polluted regions are at the top of the diagram (w.r.t. Pb, the 33 region is more polluted than the 25 region, while the opposite holds for Cd – all other regions are less polluted according to at least one of Pb and Cd). Regions that are, according to both indicators, less polluted than other regions, are placed below these more polluted regions, with a line connecting each such pair.

On the basis of such a HD, it is possible to construct a linear extension, assigning each region a unique rank, for example (12)-(35)-(21)-(25)-(58)-(11)-(57)-(33) or (12)-(58)-(35)-(11)-(21)-(57)-(33)-(25). By determining all linear extensions of a HD (Trotter, W., 1992), one obtains the rank distribution and corresponding average rank for each region. The cumulative rank frequency distributions (CRF) will uphold the monotonicity via stochastic dominance: for any 2 comparable regions, the CRF of the dominated region will lie above that of the dominating region (as can be seen in Figure 1 for regions 11 and 58). For incomparable regions, the CRF can either intersect (Figure 1, regions 25 and 57) or not (Figure 1, regions 25 and 58), in which case the regions can be thought of as a dominating and dominated region at the level of the CRF (Patil and Taillie, 2004). Incomparable regions cannot be stochastically non-monotone w.r.t. each other.

In general, among the average ranks ties occur, resulting in a weak order extension. This weak order can be further condensed by grouping together some ranks through the introduction of labels. We will investigate how to combine weak orders obtained by ranking regions according to two different layers. Two possible procedures for aggregating such data are discussed in this abstract: one based on the use of the average ranks, and one based on the use of the CRF, both in combination with the dominance-domination constraints posed by the two partially ordered sets (posets) under study.

2. Methodology

The first approach consists of computing the average ranks according to one poset, and examining to what extent these average ranks are compatible with the relations contained in another poset (stochastic domination implies that the average rank of the dominated instance will be lower than that of the dominating instance). After the necessary adjustments, the correlation between average ranks according to the first poset and the second, as well as the correlation between the average ranks according to the first poset and those same ranks adapted in order to fit the relations from the second poset are calculated.

The second approach consists in computing the CRF according to one poset, and examining to what extent these CRF are compatible with the relations contained in another poset (again keeping stochastic dominance in mind). Only after this adaptation will we compute the “average ranks” – it would no longer be accurate to call the average ranks according to one HD “average ranks” for another HD; we will therefore call these computed values “expected values” instead.

In this didactical example we will contrast the Cd and Pb pollution data for the moss layer (previously shown in Figure 1) with that of the herb layer (shown here in Figure 2). Figure 2 shows the HD for the pollution data of the herb layer and the corresponding CRF (regions 57 and 33 coincide in the CRF diagram). Remark that in the HD in Figure 1, region 11 lies above region 58, while in the HD in Figure 2, this relation is reversed. The dotted arc in Figure 2 denotes this reversal. Furthermore, even though the domination of region 58 w.r.t. region 25 in Figure 2 is not in contradiction with the HD in Figure 1 where they are incomparable, the CRF from Figure 1 do show region 25 to stochastically dominate region 58. This results in a contradiction with the relation in the second HD – hence the dashed arc denoting this inconsistency. The same holds true for regions 35 and 21. These two types of inconsistencies would become apparent through the use of expected values. The dash-dotted arc in Figure 2 however, shows an inconsistency that would not necessarily be detected by using expected values. Region 57 and 25 have intersecting CRF in the first HD, but are comparable according to the second HD. The two proposed approaches would therefore yield different results for this small data set, affecting different regions.

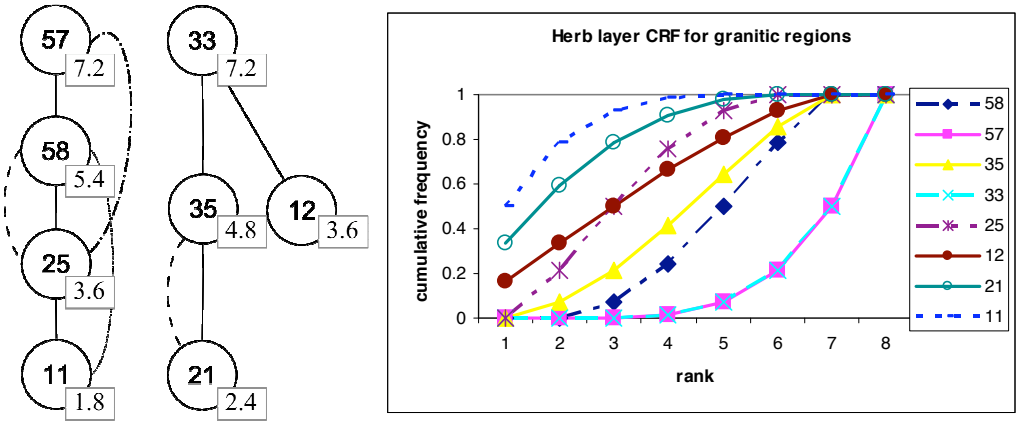

Figure 2: Herb layer HD showing average ranks (arcs are explained in the text) and CRF (regions 57 and 33 have identical CRF) for granitic regions

The first approach would collapse the expected values of regions 58 (3.0), 25 (6.4) and 11 (5.1) according to HD into a single expected value, in this case for example 5.1. In the subsequent labeling, regions 58, 25 and 11 would therefore receive the same label, as would regions 35 and 21.

The second approach would accurately show the domination of region 25 by region 57 in the second HD as being inconsistent with the CRF information from the first HD, in addition to discovering the same inconsistencies as the first approach. Instead of choosing a single expected value to assign to all pairs of conflicting regions, however, the goal is now to adapt the CRF from the first HD so that they no longer contradict the information in the second HD. We will use the Ordinal Stochastic Dominance Learner (OSDL) in order to do so. OSDL is a supervised ranking algorithm that is especially suited for dealing with distribution information (Lievens et al., in press). The algorithm is able to calculate stochastically monotone distributions on the basis of non-monotone distributions. It is exactly the distribution-based approach that leads us to prefer this algorithm to for example the Pool Adjacent Violators algorithm (Hussian et al., 2005) which would clean up the distributions by condensing them into a single rank for each region. Stochastic monotonicity does not preclude overlap between the ranges of ranks for two comparable regions. Finally, after OSDL has made the necessary adjustments to render the probability distributions stochastically monotone, the expected values can be computed and further divided into labeled classes.

Remark that the using the average ranks or CRF from one HD and making these conflict-free with the constraints posed in another HD is a directional process: regions that were incomparable in the first HD but comparable in the second, result in a new constraint on the CRF distributions (see for example regions 25 and 57 in Figure 1 and 2), possibly resulting in a non-monotone relation between the two regions. If the inverse is true, the distributions for these two regions will not need to be adapted w.r.t. each other in order to fit the second HD.

For the full data set of 59 regions the calculation of all linear extensions, needed for the rank frequency distribution, is not feasible. It is possible, however, to approximate this distribution by uniform sampling from all linear extensions via a Monte-Carlo method (De Loof et al., 2006, Bubley et al., 1999). In contrast with other approaches (Lerche and Sørensen, 2003, Lerche et al., 2003), the sampling used here is uniform, improving the accuracy of approximation. On the basis of this approximated rank frequency distribution, the same techniques as described for the example discussed in the previous paragraphs can be applied. Since the accuracy of the computed CRF is not of central importance, this should not matter much for the discussion of the results.

3. Results and Discussion

As discussed, the two described approaches to reconcile information from one HD with that contained in another, both result in new assignment of regions into labeled classes. The difference in assigned classes between these new assignments and the ones according to either of the studied HD allows for evaluation of both approaches.

Through the computation of rank probability distributions that are not in conflict with a given collection of HD, a rich source of data that respects several - possibly antagonistic - indicators is obtained. The directional approach described here opens up degrees of freedom to find a best fitting set of indicators. An interesting extension would be to allow combinations of more than two HD – a first approach could consists in simply summing and averaging the CRF of two or more different HD and relate this "averaged" CRF to another HD.

The technique is also suited to identify pairs of HD that contradict each other least. This allows minimization of the amount of data that needs to be considered by decision makers. Additionally, the experiments will also allow determination of antagonistic layers of pollution so as to better capture the diversity of the problem. Furthermore, we will discuss the advantages of the proposed approach over the alternative using only average ranks or expected values, on the basis of interpretability and rank correlation.

The proposed technique can also be interpreted as a framework for determination of partially ordered set or HD similarity, and the results can be discussed in this light as well, keeping earlier approaches in mind (Patil and Taillie, 2004, Brüggemann et al., 2003).

Bibliography

Brüggemann, R., Voigt, K., Kaune, A., Pudenz, S., Komossa, D., Friedrich, J. (1998): Vergleichende ökologische Bewertung von Regionen in Baden- Württemberg, GSF-Bericht 20/98, GSF, Neuherberg, pp. 1-148

Brüggemann, R., Pudenz, S., Voigt, K., Kaune, A., Kreimes, K. (1999):An algebraic/graphical tool to compare ecosystems with respect to their pollution. IV: Comparative regional analysis by Boolean arithmetics, Chemosphere, **38**: 2263-2279.

Brüggemann, R., Halfon, E., Welzl, G., Voigt, K., Steinberg, C. (2001): Applying the Concept of Partially Ordered Sets on the Ranking of Near-Shore Sediments by a Battery of Tests, J. Chem. Inf. Comp. Sc., **41**: 918-925

Brüggemann, R., Welzl, G., Voigt, K. (2003): Order Theoretical Tools for the Evaluation of Complex Regional Pollution Patterns, J. Chem. Inf. Comp. Sc., **43**: 1771-1779

Bubley, R., Dyer, M. (1999): Faster random generation of linear extensions, Discrete Mathematics, **20**: 81-88

De Loof, K., De Meyer, H., De Baets, B. (2006): Exploiting the lattice of ideals representation of a poset, Fundamenta Informaticae, **71**: 309-321

Hussian, M., Grimvall, A., Burdakov, O., Sysoev, O. (2005): Monotonic Regression for the Detection of Temporal Trends in Environmental Quality Data, Match - Commun. Math. Comput. Chem., **54**: 535-550.

Lerche, D., Sørensen, P.B. (2003): Evaluation of the ranking probabilities for partial orders based on random linear extensions, Chemosphere, **53**: 981-992.

Lerche, D., Sørensen, P.B., Brüggemann, R. (2003): Improved Estimation of the Ranking Probabilities in Partial Orders Using Random Linear Extensions by Approximation of the Mutual Probability, J. Chem. Inf. Comp. Sci., **53**: 1471-1480.

Patil, G.P., Taillie, C. (2004): Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization, Environmental and Ecological Statistics, **11**: 199-228.

Pudenz, S., Brüggemann, R., Komossa, D., Kreimes, K. (1998): An Algebraic / Graphical Tool to Compare Ecosystems with Respect to Their Pollution by Pb,Cd III: Comparative Regional Analysis by Applying a Similarity Index, Chemosphere, **36**: 441-450.

Lievens, S., De Baets, B., Cao-Van, K. (1992): A probabilistic framework for the design of instance-based supervised ranking algorithms in an ordinal setting, Annals of Operations Research, in press. Trotter W.T., Combinatorics and Partially Ordered Sets Dimension Theory, The Johns Hopkins University Press, Baltimore, Maryland, pp. 1-307.