ENSAI National School for Statistics and Data Analysis

Frédéric Lavancier
Professor in Applied Mathematics
ENSAI
35172 Bruz - France

Report on the habilitation thesis by Jiří Dvořák

The habilitation thesis "Nonparametric tests in spatial statistics" by Jiří Dvořák is mainly based on the following five recent articles:

[1] T. Mrkvička, J. Dvořák, J.A. González, J. Mateu (2021): Revisiting the random shift approach for testing in spatial statistics. Spatial Statistics 42, 100430.

[2] M. Ghorbani, N. Vafaei, J. Dvořák, M. Myllymäki (2021): Testing the first order separability hypothesis for spatio-temporal point patterns. Computational Statistics and Data Analysis 161, 107245.

[3] J. Dvořák, T. Mrkvička (2022): Graphical tests of independence for general distributions. Computational Statistics 37, 671–699.

[4] J. Dvořák, T. Mrkvička, J. Mateu, J.A. González (2022): Nonparametric testing of the dependence structure among points-marks-covariates in spatial point patterns. International Statistical Reviews 90, 592–621.

[5] J. Dvořák, T. Mrkvička (2022): Nonparametric testing of the covariate significance for spatial point patterns under the presence of nuisance covariates. Submitted.

As far as I can judge, the content of the thesis is original. The high percentage of coincidence detected by the Turnitin system is only due to the fact that this dissertation is a collection of reprints of several published papers co-authored by the candidate.

The above selection of articles makes a coherent whole that truly reflects the contributions and the active researches of Jiří Dvořák in the field of nonparametric tests for spatial point processes (random fields are also considered). The thesis makes a clear and pedagogical overview of these 5 papers, by explaining the statistical problems addressed in each of them, the proposed methodology and the theoretical arguments that underpin the approaches. I particularly appreciated in the report all fair comments and recommandations about the performances of each proposed method in comparison with the state of the art, not hiding the possible flaws or disappointments (if any) of the new procedures.

In a nutshell, for each statistical test considered in the above five papers, the challenge is twofold: first, suggesting a test statistic T (depending on the problem, some are simple and natural, others are more sophisticated), second, approximating the distribution of T under the null hypothesis by nonparametric Monte-Carlo simulations. To achieve this latter task, three main procedures are considered, depending on the statistical test: random permutations, random shifts, and stochastic reconstructions. Depending on the considered problem, the contribution of this habilitation thesis concerns either the first, the second or both aspects: some new test statistics are introduced, and refinements of the Monte-Carlo strategies are proposed. To be more specific, I briefly comment below on each article.

[1] This paper considers a refinement of the random shift procedure to generate replicated data under the null hypothesis H_0 of a statistical test. In this article H_0 is either independence between two random fields or independence between two random point processes. The random shift approach consists in randomly shifting the points of a point pattern so as to break any possible dependence with the other point process (and similarly for random fields). This procedure can be repeated to generate as many replications as desired, each of them being in accordance with H_0, while keeping the underlying structure of each point pattern. The main issue is that after a shift, some points may end up outside the domain of observation. To solve this issue, alternatively to the standard torus correction, it is proposed in this paper to keep only those points that remain in the domain, and to normalise (by its standard deviation) each statistical test for each replication to take into account the fact that its variance is variable, depending on the amount of deleted points. A new challenge is now to estimate this variance: several strategies are suggested, including an estimate based on the asymptotic expression of the variance, obtained for natural test statistics for the two considered testing problems. An extensive simulation study assesses the performances of the method. As a result, this approach solves the liberality problem of the torus correction in the random field case, while being more powerful than alternative procedures. In the point process case, even if the new method does not clearly outperform the torus correction, its performances are comparable. However the main advantage is that this new approach is very general, since it can be applied whatever the shape of the observation window is, unlike the torus correction that requires a rectangular window, a dramatic limitation in many applications.

[2] This article addresses the question of separability of the intensity function of a space-time point process. The main test statistic is a functional summary characteristic, constructed as the ratio between a raw estimate of the intensity and a separable estimate of the intensity, both of which depend on the spatial position u and the time t. Replications under the null hypothesis is a challenging task. The default strategy considered in this paper consists in the permutations of the time events of the space-time point pattern. While this procedure is valid for Poisson point processes, it does not provide identically distributed replications

for other point process models because permutation invariance does not hold. A simulation study clearly shows this flaw for Cox processes, when the testing procedure may become very liberal. As an alternative to permutations, replications based on stochastic reconstruction are suggested. The idea is to simulate space-time point patterns with similar prescribed statistical characteristics as the input space-time point pattern (through the minimisation of an energy function), while having a separable intensity and otherwise maximising the entropy. I found this idea promising. Unfortunately, it seems that the choice of the statistical characteristics made in the paper do not ensure the separability assumption. This is because of the inclusion of K and D_k functions that may encode the non-separability of the input process, while forcing the empirical estimate $\hat{\rho}_{sep}$ of the simulated process to be similar as the one of the input process does not guaranty that the genuine intensity of the simulated process is separable. Using the same idea, an alternative approach could have been to generate independently a purely spatial point process (with similar spatial characteristics as the input process), and a purely temporal point process (also with similar temporal characteristics as the input process), in order to finally construct a separable space-time point process by tensorial product. On the other hand, the strength of the considered testing procedure is its graphical interpretation, in line with global envelope tests widely used in spatial statistics, that allows to detect the regions (u, t) causing the rejection of the null hypothesis (if so).

[3] In this paper, a nice graphical testing procedure for independence of distributions is introduced, where the joint quantiles that deviate from the null hypothesis can be clearly identified. Given bivariate data, two test statistics are proposed: the first one is simply based on the empirical joint cdf computed at well chosen grid points, while the second one is built from the scatterplot of the joint ranks of the bivariate sample. In this latter case a kernel smoothing is subsequently applied to the scatterplot, so that the test statistic actually consists in the resulting intensity function of the joint ranks, that is expected to be constant under the null hypothesis. The replications under the null hypothesis are in turn generated by random permutations. Finally, the global envelope test machinery is applied to identify spatial regions where the joint ranks (or equivalently the joint empirical quantiles) deviate significantly from the null hypothesis, providing an insightful graphical interpretation. For the joint ranks scatterplot approach, which is by nature supported on a regular lattice, I wonder whether a direct approach could have been applied, avoiding the need to smooth the scatterplot (and thus the unpleasant necessity to choose the smoothing parameter). Many simulations are carried out that assess the performances of the tests in various situations. They show the broad generality of the approach, its effectiveness, and especially its useful graphical interpretation.

[4] This article explores in detail how to test dependencies between the locations P of a spatial point process, the marks M associated to its events, and a covariate C. Three tests are considered: a P-M test for the independence between the locations and the marks, a P-C test for the independence between the locations

and the covariate, and a PM-C test for the independence between the marks and the covariate. The first task has been already well studied in the literature and the emphasis is on the two last tests. For these ones, the test statistics are natural (the mean value of C over P for the P-C test, and the correlation between M and C computed at the locations P for the PM-C test) and the replications under the null hypothesis are carried out by the random shift approach with either the torus correction or the variance correction, as already considered in the paper [1]. This is valid for the P-C test, but as discussed in the paper, this approach is not permutation invariant for the PM-C test, so that the replications are not identically distributed in this case. To perform the variance correction, the asymptotic variance of the test statistic is obtained in both cases. For these results, the assumptions are quite strong (C is R-dependent, P is Poisson, and for the PM-C test, M and C are identically distributed). Even if we can expect the asymptotic variance to be of the same order under more general hypotheses, it would have been worthwhile to make an effort in that direction. Finally, an intensive simulation study allows for drawing relevant recommendations for the implementation of the considered statistical tests.

[5] This contribution can be seen as an alternative and an extension to the P-C test considered in [4]. In presence of a single covariate C, the test statistic consists in the Kendall correlation between C and a kernel estimation of the intensity ρ of P. In presence of several covariates $C_1, \ldots, C_m, C_{m+1}$, a nonparametric residuals is introduced, that roughly speaking compares the number of observed points in any subset B with the integral over B of $\hat{\rho}(C_1(.), \ldots, C_m(.))$, where the latter is a nonparametric estimation of ρ based on $C_1, \ldots, C_m$. Then, the Kendall correlation between C_{m+1} and these residuals is computed, leading to a partial correlation between P and C_{m+1}. This partial correlation is useful to quantify the direct dependence between C_{m+1} and P, but this statistic turns out to perform poorly in a test of independence based on the random shift approach. To overcome this flaw, a new statistic is introduced (the covariate-weighted residual mesure) that is defined as the integral over the domain of C_{m+1} with respect to the nonparametric residuals (viewed as a mesure). As illustrated in the simulation study, this clever new statistic leads to a powerful test of independence, when using the random shift approach for replications. As far as I know, this paper provides the first method to quantify and test nonparametrically the partial correlation between a point pattern and a covariate.

As it clearly appears from the above five contributions, Jiří Dvořák has a deep understanding of the statistical challenges arising for spatial and space-time point processes. He successfully achieved to introduce novel ideas on important problems, leading to powerful new methods in nonparametric tests in spatial statistics. It is worth mentioning that these contributions are systematically shared to the scientific community through public codes and packages. Nonetheless, nonparametric tests for spatial statistics is only one part of the scientific contributions of Jiří Dvořák, that also include inference for inhomogeneous Cox processes, volume estimation in stochastic geometry and depth study in multivariate datasets. These works are not discussed here.

In conclusion, based on his habilitation thesis, Jiří Dvořák has demonstrated to be a highly qualified researcher in statistics who is definitively qualified to supervise research and be an associate professor in Applied Mathematics.

Rennes, May 16, 2023.

Frédéric Lavancier.