Spatial clustering analysis using copulas and point patterns

M. Omidi[1], C. Ayyad[2,*], J. Mateu[2], M. Mohammadzadeh[1], I. Tamayo[3]

[1] *Department of Statistics, Tarbiat Modares University, Tehran, Iran; mehdi.omidi@modares.ac.ir, mohsen_m@modares.ac.ir*
[2] *Department of Mathematics, University Jaume I, Castellón, Spain; ayyad@uji.es, mateu@uji.es*
[3] *Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain; ibontama@gmail.com.*
[*] *Corresponding author*

Abstract. *Rat sightings can be described by spatial coordinates in a particular region of interest defining a spatial point pattern. In this paper we investigate the spatial structure of rat sightings and its relation to a number of distance-based covariates that relate to the proliferation of rats. We use copula functions to build a particular spatial multivariate distribution using univariate margins coming from the covariate information. We use maximum likelihood together with the Bee algorithm to estimate the corresponding parameters, and perform prediction of rat sightings according to the predefined six focuses in Latina (Madrid).*

Keywords. *Bee algorithm, Copula functions, Rat sightings, Spatial copula, Spatial point patterns.*

---

1 Extended Abstract

In the last few decades, changes in cities have facilitated the proliferation of pests and corresponding diseases associated with them. Cities have expanded through natural habitats of rodents and other pests, resulting in the reactivation of diseases that were thought to be extinct. Urbane plagues are often the cause of important expenses of public administrations in tasks and strategies trying to eradicate them. One of the most harmful plagues comes through the *Rattus norvegicus*, prevalent species in the majority of European cities.

Under an accumulation of favourable situations (Ayyad *et al.*, 2014), such as presence of water, green areas, markets and cat feeding stations, the *Rattus norvegicus* proliferation takes place. We identified up to six focuses along the region based on a number of markets and water sources. The locations of the focuses were selected so that they favour the presence and accumulation of rats, and cover most of the

region of interest.

When a sighting of a rat is reported to the Technical Unit for Vector Control (TUVC), information about the location, date and person reporting that sighting is collected and entered in a dedicated database. Each reported pest sighting corresponds to an individual record in this database. Our data contain the locations of 470 validated rat sightings reported to the TUVC from January 2002 to December 2008. Rat sightings and all 8985 buildings were geo-referenced and mapped in Latina district (Madrid, Spain), as shown in Figure 1.

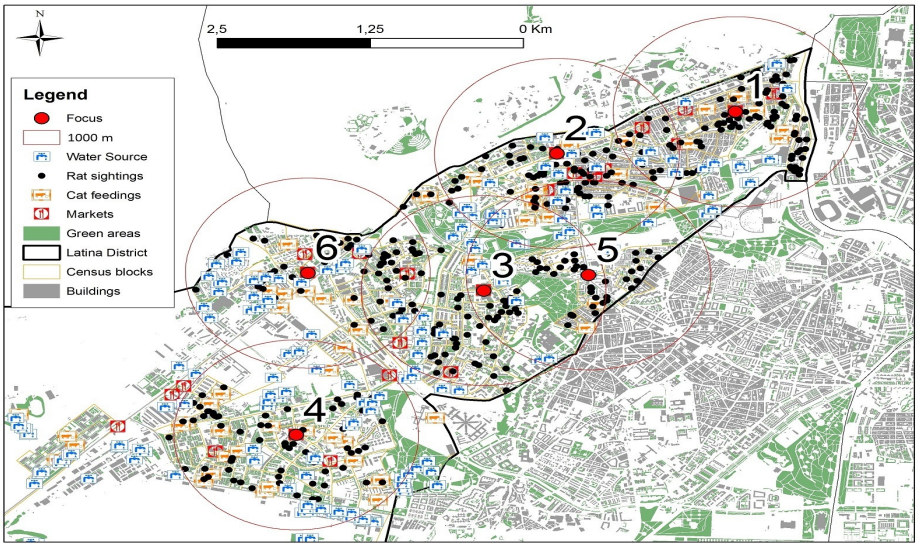

Figure 1: Latina district with indication of the locations of the 470 rat sightings (black dots), the potential focuses (red dots), the 8589 buildings (grey dots), and the 205 census blocks in which the region has been divided (yellow rectangles)

We also considered the following additional variables playing a role in our statistical approach: (a) **Distance to focuses**, considering buffers around each focus of 1000 meters; (b) **Minimum distance to nearest covariates**, and (c) **Angular direction** of the rat location with respect to the location of the focus.

We considered four main covariate information in terms of distances to water sources, to green zones, to markets, and to cat feeding stations. These distance-based variables are represented by D_{Ws}, D_{Gz}, D_M and D_{Cf}. But we are interested in modeling the distribution of the minimum distance to nearest covariates, i.e. of $Z = \min(D_{Ws}, D_{Gz}, D_M, D_{Cf})$. We then first fitted the Weibull distribution to any individual covariate, and then calculated the minimum value of such covariates $z = \min(d_{Ws}, \ldots, d_{Cf})$ with distribution $F_Z(z) = 1-(1-F_{D_{Ws}}(z))(1-F_{D_{Gz}}(z))(1-F_{D_M}(z))(1-F_{D_{Cf}}(z))$, with $z \geq 0$. In particular, Figure 2 depicts clockwise-type orientation to measure the three important elements: distance to focus (D_F), direction (Φ), and the minimum value amongst the nearest covariates (Z) associated to each rat location.

Based on the highly structured behavior of rats, it is necessary to use a powerful tool for modeling the spatial dependency structure of the rat sightings. Copulas (Nelsen, 2006) are multivariate distributions with uniform margins which provide a tool to describe the dependency structure among variables. In recent years, spatial copulas are widely applied and developed (see, for example, Bardossy and Li (2008),

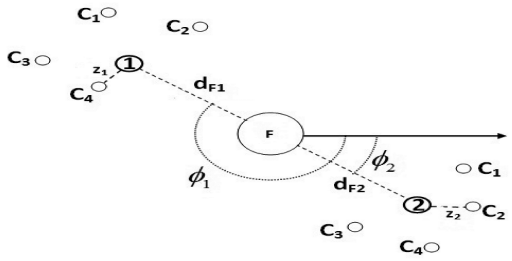

Figure 2: Calculation of distances and orientations (clockwise type) with respect to focus F.

			C_2^{Cl}			
F		C_1^{Cl}	θ	β	C^F	C^{GH}
F1	Est.	0.1195	0.05995	1.0920	0.8806	1.1177
	$-AIC$	-0.6732	1.5586		**6.5332**	5.2334
F2	Est.	0.3179	0.1830	1.0748	1.2489	1.1236
	$-AIC$	6.6349	4.972		**9.5414**	5.3352
F3	Est.	0.2311	0.2525	1.009	1.0055	1.0650
	$-AIC$	7.0298	6.2214		**11.2416**	0.4698
F4	Est.	0.1264	0.1251	1.0028	0.4440	1.0274
	AIC	**0.4662**	2.4053		0.6752	1.6120
F5	Est.	0.0007	0.0011	1.0012	0.0012	1.0015
	AIC	2.0351	4.0036		**0.6752**	1.6120
F6	Est.	0.0011	0.0097	1.0034	0.0016	1.0009
	AIC	2.0360	4.0274		**2.0122**	2.0906

Table 1: Parameter estimates for the selection of copula families, and the AIC values.

and Kazianka and Pilz (2010)). For the analysis of spatial point pattern data, there is just one work by Kuethe *et al.* (2009) who applied the bivariate t-copula to model the spatial point pattern of housing price values in an urban area. In this paper we consider a trivariate copula to explore the spatial dependency of the data. In particular, we consider four trivariate copulas, Clayton with $\theta > 0$ (C_1^{Cl}), Clayton 2-parameters with $\theta > 0$ and $\beta \geq 1$ (C_2^{Cl}), Frank with $\theta > 0$ (C^F), Gumbel-Hougaard with $\theta > 1$ (C^{GH}) as follows:

$$\begin{aligned}
C_1^{Cl}(u_1,u_2,u_3,\theta) &= (u_1^{-\theta}+u_2^{-\theta}+u_3^{-\theta}+-2)^{-\frac{1}{\theta}} \\
C_2^{Cl}(u_1,u_2,u_3,\theta) &= \{[(u_1^{-\theta}-1)^{\beta}+(u_2^{-\theta}-1)^{\beta}+(u_3^{-\theta}-1)^{\beta}+1]^{\frac{1}{\beta}}\}^{-\frac{1}{\theta}} \\
C^{F}(u_1,u_2,u_3,\theta) &= -\frac{1}{\theta}\log\{1+\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)(e^{-\theta u_3}-1)}{(e^{-\theta}-1)^2}\} \\
C^{GH}(u_1,u_2,u_3,\theta) &= \exp\{-[(-\log u_1)^{-\theta}+(-\log u_2)^{-\theta}+(-\log u_3)^{-\theta}]^{\frac{1}{\theta}}\}
\end{aligned}$$

Note that any of these families cover $M(\cdot)$ and $\Pi(\cdot)$ for especial values of their parameters (Nelsen, 2006). We used the IFM procedure to estimate the copula parameters through maximum likelihood. To minimize the -log-likelihood function, $-\log L_c(\theta|F_{D_F},F_\Phi,F_Z) = -\sum_{i=1}^{n}\log c(F_{D_F}(d_i),F_\Phi(\phi_i),F_Z(z_i),\theta)$, where c refers to the density of the trivariate copula, the Bees Algorithm (BA) (Pham *et al.*, 2005) was used. The results are shown in Table 1 which contains the estimates of the copula parameters, and the AIC for each focus. We note that Frank copula is the best function fitting the rat sighting spatial dependency structure for all focuses except for Focus 4, for which the Clayton family with one parameter shows the best fit. Moreover, the values of copula parameters in focuses 5 and 6 show that the corresponding variables are close to being independent.

If we are interested in predicting the distance of rat sightings to focus or the corresponding direction,

	Distance to F (meters)				Direction (degrees)			
		Predicted				Predicted		
	Obs.		%95 CI		Obs.		%95 CI	
F	Mean	Mean	Lcl	Ucl	Mean	Mean	Lcl	Ucl
F1	514.206	518.333	506.585	530.081	226.967	226.757	223.894	229.621
F2	466.382	464.826	452.721	476.931	124.411	123.267	119.720	126.814
F3	635.452	636.837	626.932	646.742	165.991	165.853	163.948	171.758
F4	571.351	575.897	569.836	581.957	1281.188	181.911	178.967	184.854
F5	579.407	586.899	586.886	586.912	165.346	165.349	165.342	165.356
F6	603.280	671.530	671.489	671.572	184.588	184.585	184.580	184.590

Table 2: Mean and 95% CI for prediction for variables D_F and Φ for each focus.

we can also obtain a close form of the conditional distribution, where the values of the observed mean, the predicted mean and the 95% confidence interval (CI) for mean prediction are shown in Table 2. Because of the independent structure in focuses 5 and 6, the conditional mean and CI for prediction of the distance and direction both tend to the mean of the Weibull distribution $(\beta\Gamma(1+\frac{1}{\alpha}))$ and the mean of the Normal distribution, as expected.

Acknowledgments. Work partially funded by grant MTM2010-14961 from the Spanish Ministry of Science and Education.

References

[1] Ayyad, C., Mateu, J. and Tamayo, I. (2014). Spatial modelling of rat sightings in relation to urban multi-source focus. Submitted.

[2] Bardossy, A. and Li, J. (2008). Geostatistical Interpolation Using Copulas. *Water Resources Research*, **44**, 44:W07412.

[3] Kazianka, H. and Pilz, J. (2010). *Spatial Interpolation Using Copula-based Geostatistical models*. Springer, Berlin, 307-320.

[4] Kuethe, T.H., Hubbs, T. and Waldorf, B. (2009). Copula Models for Spatial Point Patterns and Processes. Third World Conference of Spatial Econometrics, Barcelona, Spain.

[5] Nelsen, R.B. (2006). *An Introduction to Copulas*, Springer.

[6] Pham, D.T, Ghanbarzadeh, A., Koc, E., Otri, S., Rahim, S. and Zaidi, M. (2005). The Bees Algorithm. Technical Note, Manufacturing Engineering Centre, Cardiff University, UK.