Ann Oper Res
DOI 10.1007/s10479-011-0841-3

Multiobjective evolutionary algorithms to identify highly autocorrelated areas: the case of spatial distribution in financially compromised farms

Carlos R. García-Alonso · Leonor M. Pérez-Naranjo · Juan C. Fernández-Caballero

Abstract Local Indicators of Spatial Aggregation (LISA) can be used as objectives in a multicriteria framework when highly autocorrelated areas (hot-spots) must be identified and geographically located in complex areas. To do so, a Multi-Objective Evolutionary Algorithm (MOEA) based on SPEA2 (Strength Pareto Evolutionary Algorithm v.2) has been designed to evaluate three different fitness functions (fine-grained strength, the weighted sum of objectives and fuzzy evaluation of weighted objectives) and three LISA methods. MOEA makes it possible to achieve a compromise between spatial econometric methods as it highlights areas where a specific phenomenon shows significantly high autocorrelation. The spatial distribution of financially compromised olive-tree farms in Andalusia (Spain) was selected for analysis and two fuzzy hot-spots were statistically identified and spatially located. Hot-spots can be considered to be spatial fuzzy sets where the spatial units have a membership degree that can also be calculated.

Keywords Multiobjective evolutionary algorithms · Spatial analysis · Local indicators of spatial aggregation · Fuzzy hot-spots · Financially compromised areas

C.R. García-Alonso (✉)
ETEA, Business Administration Faculty, University of Córdoba (SPAIN), Escritor Castilla Aguayo, 4, 14004 Córdoba, Spain
e-mail: cgarcia@etea.com

L.M. Pérez-Naranjo
Business Administration Department, University Pablo de Olavide, Carretera de Utrera km 1, 41013 Sevilla, Spain
e-mail: leonorpn@hotmail.com

J.C. Fernández-Caballero
Computer Science and Numeric Analysis Department, University of Córdoba, Rabanales Campus, Building Einstein, 14071 Córdoba, Spain
e-mail: fernandezcaballero@gmail.com

Published online: 20 January 2011

1 Introduction

The identification of spatial units (SU) where a phenomenon is significantly concentrated is specially relevant in many decisional situations both to identify the causes of such spatial aggregation and to support decision making (Baddeley et al. 2006; Andrienko and Andrienko 2006; Gutiérrez et al. 2008; Noyan 2010). According to Agarwal et al. (2005) and Moreno et al. (2008), hot-spots are those spatial conglomerates where the geographical distribution of the phenomenon is highly and significantly autocorrelated.

Spatial analysis methods are widely used to identify hot-spots and geographically locate them. Holloway et al. (2007) grouped them into: (i) those methods that specifically use spatial econometrics and (ii) those that apply any Geographical Information System (GIS) tool. In the first group, there are many methodological approaches from classical Poisson analysis of random occurrences in space (Youssef et al. 1991) to artificial neural networks, evolutionary algorithms and fuzzy logic (Fischer 2006; Coello-Coello et al. 2007); from among these some well known techniques are Local Indicators of Spatial Aggregation—LISA (Ord and Getis 1995; Anselin 2002; Holloway et al. 2007). GIS utilities to analyse geo-referenced spatial data also include quite a number of possibilities like: Kernel estimation (Silverman 1986), distance decay estimates (Staal et al. 2002) and dynamic Minimum Spanning Trees (Assunçao et al. 2006). All of them show smooth surfaces in maps that highlight relevant areas.

Instead of analysing global structures and tendencies in the spatial distribution of a variable, LISA methods statistically contrast the existence of local concentrations (autocorrelation) of significantly higher or lower variable values compared to its mean: hot-spots. These methods are based on standard correlation and Durbin-Watson statistics in a time-series context and try to detect spatial concentrations defined by a local dependence phenomenon. Some examples of standard LISA methods are: local Moran's I, local Geary's C and Getis and Ord's G (Anselin 1995; Ord and Getis 1995). All of them are useful but their results must be interpreted carefully because of their different ways of approximating the spatial distribution of the phenomenon under study. For example, Moran's I coefficient can be interpreted as a covariance, the calculation procedure for Geary's C is similar to a variance analysis where the influence of variable values is greater than in Moran's I (and its interpretation is the opposite as well) and, finally, Getis and Ord's G is an association measurement.

Each spatial econometric method, all of them useful, provides different sizes, shapes and locations of hot-spots for the phenomenon under study (Moreno et al. 2008). Authors usually focus their attention on only one method to explain the spatial distribution of a specific phenomenon but it is difficult to find papers that simultaneously combine various methods to obtain generalized conclusions about their potential agreement rates. Some GIS utilities, like map algebra, can be used to combine spatial projections obtained from econometric models (Pérez-Naranjo and García-Alonso 2005; Moreno et al. 2008), but this procedure is almost always subjective. Gonçalves and Estellita (2002) used standard linear programming in spatial analysis to face the multi-objective problem generated when different spatial approaches are considered, but unfortunately this problem is not linear (Hof and Bevers 2000).

Spatial analysis needs a wide enough geographical distribution of geo-referenced data because it is based on the geographical contiguity of SU. The existence of unknown or dark zones, where analysts cannot assign any data, is very frequent in complex areas. This problem can be solved substituting missing data with data predicted by statistical models (De Pinto and Nelson 2007) or by spatial interpolation (Andrienko and Andrienko 2006). When the statistical distribution of the phenomenon is unknown and a spatial pattern exists according to empirical evidence, like in the agrarian sector, spatial interpolation is statistically more efficient than ignoring these missing values (Griffith 2003).

Isolated SU with high original values and autocorrelation scores are very easy to identify using, for example, GIS utilities. Nevertheless, the identification and the geographical location of hot-spots is a challenge to be dealt with because their identification and geographical location in large and heterogeneous areas with dark zones is a non-supervised clustering one. In this framework, each spatial econometric model is a non-linear objective to be optimized in search of a compromise (García-Alonso and Pérez-Naranjo 2007). This model must handle unknown algebraic and spatial trade-offs.

A multi-objective problem is a decisional situation where the decision maker (DM) has to optimize some objectives with unknown trade-offs between them. When non-linear functions have to be optimized in a potential non-convex decisional space, Multi-Objective Evolutionary Algorithms (MOEA) can be used (Bäck 1996; Kalyanmoy 2004; Tavares-Pereira et al. 2007) to successfully approximate the Pareto set instead of other traditional techniques like the preference-based approach and Monte-Carlo simulation (Coello-Coello et al. 2007). Rather than performing mathematical optimization techniques on the objectives, a MOEA iteratively tries to obtain better and better solutions in the variable space while avoiding local optima. These solutions are evaluated using a fitness function which derives from the objective set (Bäck 1996), for example: weighted sum of objectives, rank dominance, Pareto ranking, strength value based on dominance, fuzzy inference and so on (Coello-Coello et al. 2007). Each solution set or population is improved in a step-by-step iterative process, generation by generation, using evolutionary operators: selection, crossover and mutation. Once a stopping procedure is fulfilled, MOEA results approximate the Pareto set of efficient solutions. According to DM needs, MOEA can use different fitness types for the same problem to evaluate and compare alternative Pareto sets.

According to Coello-Coello et al. (2007) there are three categories of MOEA techniques: "a priori", progressive and "a posteriori". In the first one, the DM has to define the relative relevance of each objective before solving the resulting model. This characteristic is the major drawback of these algorithms (Wilson and MacLeod 1993; Das and Dennis 1997) because the selection of the objective weights is subjective and, depending on it, some relevant solutions can be missed since the variable space is arbitrarily limited. In progressive techniques, DM incorporates his preferences in the process in an interactive way (Barbosa and Barreto 2001). Obviously, DMs have to be available to evaluate how appropriate a specific solution set really is. DM subjective preferences again limit the search space and, independently of their applicability, an interactive process can be used whatever the "a priori" or "a posteriori" technique that has been selected. Finally, "a posteriori" techniques (Das and Dennis 1997; Srigiriraju 2000; Laumanns et al. 2006) explore the whole variable space trying to obtain, theoretically, the Pareto set or, in practice, as many elements within it as possible.

When spatial analysis methods are used as objectives, all of them have an identical relevance "a priori". In this specific situation "a posteriori" techniques are the most appropriate. These techniques group many different MOEA strategies: independent sampling (Srigiriraju 2000), criterion selection—Vector Evaluated Genetic Algorithm (VEGA) and extensions (Kursawe 1991; Hajela and Lin 1992), aggregation selection—weighted sums, constraint and objective combinations or hybrid search approaches (Loughlin and Ranjithan 1997; Ishibuchi and Murata 1998; Deb 1999), ϵ-Constraint (Laumanns et al. 2006), Pareto sampling techniques (Coello-Coello et al. 2007) and so on. Whatever technique is selected, the scalability of the problem is an issue to be dealt with, and there is no "a priori" best MOEA strategy for any one problem (Wolpert and Macready 1997).

Our MOEA is based on the Strength Pareto Evolutionary Algorithm v.2 (SPEA2) which has been considered as an "a posteriori" approach to multiobjective optimization (Zitzler and

Thiele 1999; Zitzler et al. 2001). This algorithm is very flexible because it was specifically designed to integrate different MOEA strategies based on specific fitness functions (Coello-Coello et al. 2007). One of its main characteristics is that it needs an external file (ENDSF) to accumulate the non-dominated solutions obtained in each generation. The "strength" of each non-dominated solution is evaluated proportionally, generation by generation, taking into account the number of solutions that it dominates (dominance count) and the number of individuals that dominate it (dominance rank). This strategy is called fitness-grained assignment based on dominance and clustering and it combines both single dominances in a unique dominance ranking. Other strategies are dominance rank (Horn et al. 1994; Zydallis et al. 2001) and dominance depth (Deb et al. 2002), nevertheless, our strategy seems to be more efficient for identifying hot-spots in space. In order to guide the algorithm, SPEA2 uses a nearest neighbour density estimation technique and, to preserve extremely efficient solutions, it truncates the external file thus removing unnecessary elements.

The objective of this paper is to demonstrate that MOEA, based on the SPEA2 strategy, is an interesting methodological approach which can be used to identify and geographically locate highly autocorrelated zones (hot-spots) by combining k econometric methods and different fitness functions even when there is incomplete spatial data and, due to that, spatial interpolation is needed. The spatial distribution of the financial risk of Andalusian olive-tree farms in dry farming at municipality level was selected to be analysed. This was because the identification of the financially compromised olive-growing-areas is strategically relevant because olive trees are the most important crop in this region (southern region in Spain).

This paper is structured as follows: in Sect. 2 the methodology is described; Sect. 3 presents the experimental design; results from MOEA models are statistically described and analysed in Sect. 4; and, finally, some illustrative comments and conclusions are drawn in Sect. 5.

2 Methodology

In order to identify hot-spots, the SPEA2 algorithm has been used as the primary MOEA strategy and Moran's I, Geary's C and Getis and Ord's G were the spatial autocorrelation methods selected as objectives to approximate the spatial distribution of the phenomenon. Results obtained from the standard SPEA2 fitness function were compared to those obtained using both the weighted sum of objectives (Ishbuchi and Murata 1996; Coello-Coello et al. 2007) and a fuzzy evaluation of weighted objectives (Lee and Esbensen 1997; Wang and Terpenny 2005). This procedure is used to check the hypothesis that hot-spots are fuzzy and the membership degree of each of their potential SU can be approximated.

2.1 Univariate spatial analysis

Hot-spots group an unknown number of spatially close SU with optimum and uniform LISA scores that can be estimated by their means and standard deviation (SD). Since the municipalities are the highest precision SU in the area under study, they were selected to determine and geo-reference LISA scores. So, each SU_i has an associated LISA scores vector $[I_i, C_i, G_i]$ calculated according to the value of the variable selected in it and in its nearest neighbourhoods.

In regions where SU sizes are very heterogeneous, the use of real or Euclidean distances can bias the real spatial distribution of the phenomenon (Anselin 2002; Pérez-Naranjo and García-Alonso 2005). This fact is due to the standard contiguity matrix structure that considers that a common SU boundary defines spatial neighbourhoods. This assumption is correct

when SU have similar sizes, but it is problematic when very small and very big SU co-exist along complicated boundaries.

In order to carry out LISA methods in complex regions, the neighbourhood definition can be reformulated using a distance criterion (Emir-Farinas and Francis 2005). According to that, neighbourhoods are considered to be both all the SU with a boundary in common and also those SU within a statistically defined radius (Anselin 2002). In a first step, real neighbourhoods with boundaries in common of $SU_j \in M$ (M being the SU set) are included in NH_j (neighbourhood set of SU_j). Other SU_j neighbourhoods are sequentially identified and included in NH_j following the procedure: (i) the nearest neighbourhood $SU_i \notin NH_j$ to SU_j is identified, (ii) the mean square error (MSE) of the distances between SU_j and all its neighbourhoods, including SU_i, in NH_j is calculated and (iii) if the MSE is lower than a predefined value, SU_i is included in NH_j and the process continues in (i); if not, it stops and SU_i is not included in NH_j.

2.2 The multiobjective evolutionary algorithm

Taking into account the framework under study, one solution is a SU set $(SU_1, SU_2, \ldots, SU_n)$ selected from the set of all the possible combinations without repetition of n SU, and it is a potential hot-spot. For a specific solution, the means of its LISA scores are estimators of its global autocorrelation relevance, their corresponding SD are estimators of the spatial uniformity of LISA scores and the vector $(\bar{I}, SD_I, \bar{C}, SD_C, \bar{G}, SD_G)$ can be determined easily. From a theoretical point of view, a hot-spot can be identified through: (i) optimizing the means $(Max\bar{I}, Min\bar{C}, Max\bar{G})$ and (ii) minimizing the SD $(MinSD_I, MinSD_C, MinSD_G)$. In addition to the $2k$ LISA objectives (k being the number of econometric methods), another objective is necessary to obtain spatially close SU sets: the minimum path that joins all the SU in the solution also has to be minimized. In the end, $2k + 1$ objectives must be analysed.

Based on the standard SPEA2 strategy, some improvements have been developed to identify hot-spots in space (Table 1 shows the algorithm structure). These improvements begin with the initial population design where SU can be selected at random in both the complete region under study or in a pre-classified region. Taking geographical diversity into account, an additional loop has also been included to analyse more than only one initial solution. Finally, three different fitness functions are used to guide the algorithm: the standard SPEA2 fitness function, the weighted sum of objectives (weights selected at random) and, finally, a fuzzy inference engine that evaluates weighted objectives.

In order to minimize the minimum paths (*MinPath*), a relative distance must be defined to avoid undesirable bias. Therefore, $\forall j, i$ SU_j to SU_i distances d_{ji} have been substituted by a relative proximity rate rp_{ji} in $[1, rp_{\max}] \subset Q$ (see another illustrative example in Hodgson and Jacobsen 2009). When SU_i is a real SU_j neighbourhood, then $rp_{ji} = 1$. Once this first relative proximity rate is assigned, the procedure identifies the neighbourhoods that are not part of those SU that belong to this first level ($rp_{ji} = 1$) and assigns them the relative proximity rate $rp_{ji} = 2$. Sequentially, neighbourhoods of later levels identify the following relative proximity rate until a predefined maximum $rp_{\max}$ is reached (Fig. 1 shows an illustrative example). This procedure minimizes the effects of SU size in LISA estimation and in MOEA.

The fitness function based on the fuzzy evaluation of weighted objectives is not standard (Wang and Terpenny 2005). All membership functions (MF)—semantic labels like HIGH and LOW (1)—were designed as symmetric triangular or Z functions. Let m be the number of MF for each of the $2k + 1$ objectives selected, the total number of fuzzy rules like (1) that the inference engine must analyse for each pair of MF, MF_h^-—at the left—and MF_h^+—at the

Table 1 Structure of the MOEA algorithm for the evaluation of hot-spots in space (based on SPEA2 model, Zitzler et al. 2001). Three different fitness functions have been included and evaluated

procedure based on SPEA2 (Population size N', number of generations g, kth objective $f_k(\boldsymbol{x})$)
  Create the external non dominated solution file (*ENDSF*)
  For $i = 1$ to s **do** ▷ *s being number of different initial solutions.*
    Initial population design, S ▷ *At random. Based or not on regional divisions.*
    for $i = 1$ to g **do**
      Compute the fitness of each individual in S and *ENDSF* ▷ *Standard SPEA2, weighted sum of objectives and/or fuzzy evaluation of weighted objectives.*
      Preserve all non-dominated solutions in *ENDSF*
      If *ENDSF* is too big then the truncated operator removes unnecessary solutions ▷ *Extreme solutions are preserved in the variable space.*
      Empty registers in *ENDSF* are filled out using dominated solutions
      Binary tournament selection with replacement ▷ *Elitism is optional. Two stopping criteria (MCE) are analysed.*
      Crossover ▷ *Elitism is optional. Crossover can be simple or double.*
      Mutation ▷ *Elitism is optional. At random, distance-based and fitness-based.*
      Repair process ▷ *Structural and technical infeasibilities.*
    end for
  end for
end procedure

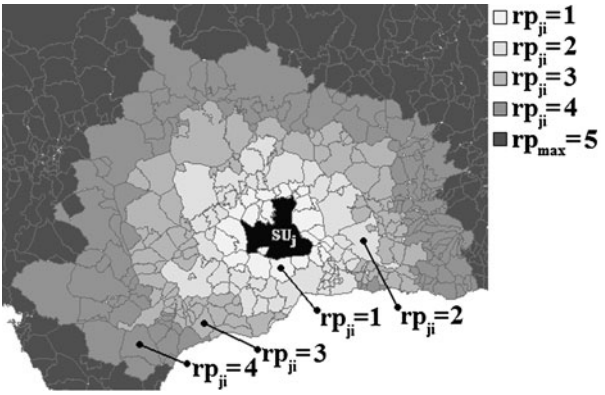

Fig. 1 Example of the determination of relative proximity rates (*grey scale*) for a specific SU_j (*black*) and $rp_{\max} = 5$ (*darker grey*)

right—($h = 1, 2, \ldots, 2k + 1$), determined by each real objective value is $2^{(2k+1)}$. Each objective value in the vector ($\bar{I}, SD_I, \bar{C}, SD_C, \bar{G}, SD_G, MinPath$) determines MF_h^- and MF_h^+, 2^7 being the total number of rules like (1) that the inference engine must evaluate to determine the final fitness value.

IF
  $\bar{I}$ is HIGH and SD_I is LOW and $\bar{C}$ is LOW and SD_C is LOW and $\bar{G}$ is HIGH and SD_G is LOW and the Minimum Path is *LOW*
THEN
  Fitness value is HIGH (1)

Obviously there is no expert knowledge to determine the resulting MF of the fitness for each 2^{2k+1} combination. For each solution and w_h being the weight of objective h (selected at random for each initial solution), the MF of the fitness is determined by:

$$MF = \sum_{h=1}^{2k+1} w_h MF_h^{-/+} \Big/ \sum_{h=1}^{2k+1} w_h \quad \text{for each } 2^{2k+1} \text{ combination.} \tag{2}$$

Usually, but not always, $\sum_{h=1}^{2k+1} w_h = 1$. The appropriate MF for each objective value is selected according to its orientation: maximization or minimization. Once the fitness MFs for each 2^{2k+1} combination have been determined, Mizumoto's Product-Sum-Gravity method with composition (Asai 1995) is used to calculate the fitness value of the solution.

A standard tournament with replacement of two solutions has been selected for the MOEA selection operator. During this process, both the weak and strict Pareto optimality criteria (Coello-Coello and Lamont 2004) are used. Dominated solutions can be substituted at random by completely new feasible solutions to increase diversity in the variable space. Finally, the selection operator includes two different stopping procedures based on the MSE. The first one evaluates the algorithm aptitude, estimated by the fitness MSE, which was calculated taking into account the best efficient solutions obtained in each generation (discarding the first ones). The second stopping procedure explores the whole ENDSF calculating a global fitness MSE. When one of them is repetitively below a predefined value or the maximum number of generations is reached, the algorithm stops and the solutions are saved in the ENDSF.

Crossover can be simple (solutions are cut in the same position between two SU) or double (there are two cuts in each solution) but the procedure is standard. A mutation operator replaces one or more SU in a parent solution with other different SU. Mutation is performed at random or can be distance-based or fitness-based (both selected at random). The last two mutation types greatly increase the intensity of the search and must be carefully adjusted to avoid premature convergence to a local optimum. Once a SU in the solution is selected to be mutated, distance-based mutation identifies the SU that minimizes the minimum path to the remaining SU. On the other hand, fitness-based mutation selects the SU that maximizes the fitness function. The elitism of the best solution obtained is always optional.

Rather than using a strategy based on penalty functions to manage non-feasible solutions (Coello-Coello et al. 2007), our MOEA includes a repair process (Michalewicz and Schoenauer 1996; Smith and Coit 1997). In our framework, the second choice is very efficient because the mutation operator (at random, distance-based or fitness-based) can be used to reach a new feasible solution similar to the original one. There are two different types of infeasibility: structural and technical. The first case is very simple: it is absolutely impossible to have repeated SU in a solution. This circumstance can sometimes occur when the crossover operator is performed. The most relevant technical infeasibility is due to the minimum path control, that is, the minimum path that links all the SU in a solution should be lower than or equal to a predefined value. Due to Pareto optimality definitions, the variable space can include solutions with a very large minimum path in the space of objectives. In our problem domain, it is impossible to consider efficient solutions like that (dispersed hot-spots) and, therefore, the minimum path has to be constrained.

2.3 Identification of hot-spots in space

At the end of the procedure, the ENDSF stores the potential hot-spots. Each individual SU there is analysed to determine how many times it appears and the results are plotted in a

standard histogram. Extreme SU, at the right hand side of the histogram, identify those SU that appear an extremely high number of times in potential hot-spots. These extreme SU are identified using standard Q-Q Plots (Beirlant et al. 2004) considering that extremes can be adjusted to an exponential probability distribution. These extreme SU finally identify hot-spots and can be geographically located on a map using a GIS.

According to this procedure, many SU in the ENDSF are ignored. Taking this fact into account, the original LISA values of the SU identified as hot-spots are substituted by the corresponding LISA means, calculated without them, and the MOEA is run again to identify and locate new hot-spots. This procedure does not penalise primary hot-spot neighbourhoods and, therefore, they can appear in the new hot-spot set.

3 Experimental design

3.1 Databases: Andalusian olive tree farms in dry farming

Andalusia is one of the largest regions in Spain as well as in the European Union. It is located in the south of Spain, limited at the south by the Mediterranean Sea and the Atlantic Ocean. One of its main economic sectors from both social and territorial points of view is agriculture and olive trees in dry farming are its most important crop (Junta de Andalucía 2004).

In 2003 a complex survey was carried out to evaluate the socio-economic structure of Andalusian farms in the agricultural year 2000 and 252 dry farming olive-tree farmers were interviewed in 80 municipalities (SU). Socioeconomic variables (costs, revenues, hand labour and so on) were simulated 2619 times in each agricultural year from 2001 to 2015. From a sustainability point of view, the net margin, obtained by subtracting total direct costs from total revenues including subsidies, is the most important variable to evaluate the financial risk of a specific farm or territory. For each SU in the agricultural year 2001, the financial risk was calculated as the probability of having a net margin below or equal to zero.

Andalusia has 770 SU but only 80 were surveyed and, consequently, many of them have missing values. Taking into account that olive-tree farms are spread throughout most of the region and that cultivation maps are available, spatial interpolation is feasible. Let $M \in Q^p$ be the set of SU in the zone under study and $S \subset M$ the set of municipalities where nf_i farms have been surveyed ($i = 1, 2, \ldots, s$; s being the number of municipalities with almost one surveyed farm) and where a financial risk value fr_i was really calculated. For each municipality $j \notin S$, let $NH_j \subset M$ be the set of its nearest neighbourhoods h with a relative proximity rate (Sect. 2.1) $rp_{jh} \leq rp_{\max}$, $rp_{\max}$ being a predefined value. When the number of municipalities n_j with surveyed farms in NH_j is greater than a parameter $nb_{\min}$, the estimated/interpolated financial risk efr_j is calculated by:

$$efr_j = \sum_{i=1}^{n_j} (nf_i \times fr_i / d_{ji}) \Big/ \sum_{i=1}^{n_j} (nf_i / d_{ji}) \quad \text{when } n_j \geq nb_{\min} \; \forall j \in M \text{ not in } S \tag{3}$$

where d_{ji} is the distance between j and its nearest neighbourhood $i \in S$ defined by the relative proximity rate. This approach improves the standard spatial interpolation used by Andrienko and Andrienko (2006) when there are a variable number of real observations in each SU. For $rp_{\max} = 4$ and $nb_{\min} = 3$, the resulting number of interpolated SU was 461. Figure 2a shows the spatial distribution of financial risk of olive-tree farms in dry farming

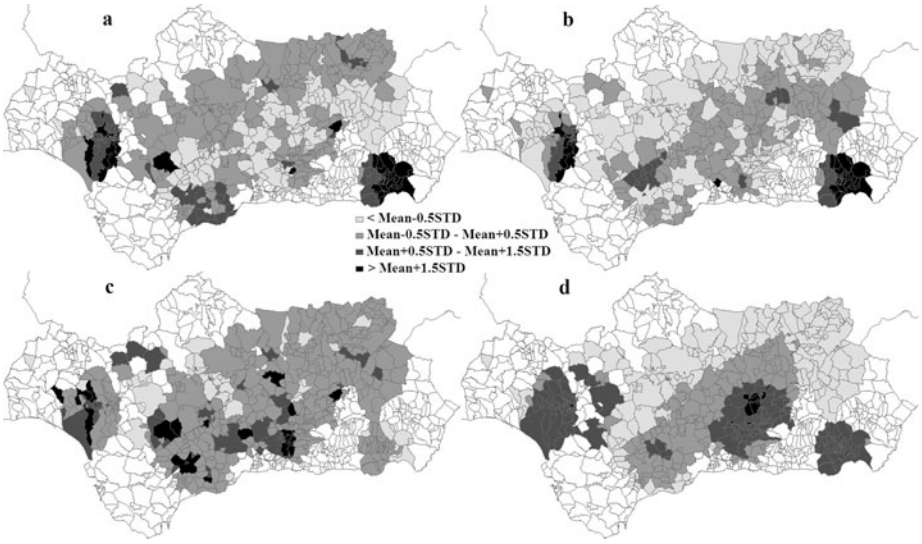

Fig. 2 Spatial distribution of financial risk—real and interpolated values (**a**), rescaled Moran's I (**b**), rescaled Geary's C (**c**) and rescaled Getis and Ord's G (**d**)

Table 2 Basic statistics for both original (80 municipalities) and interpolated (461 municipalities) financial risks and local autocorrelation scores (Moran's I, Geary's C and Getis and Ord's G)

	Financial risk		Local autocorrelation methods (LISA)[a]		
	Original	Interpolated	I	C	G
Average	0.4117	0.4485	0.3648	0.1372	0.5585
SD[b]	0.2645	0.1901	0.1531	0.0760	0.1837
Median	0.4053	0.4143	0.2958	0.1127	0.5377
Minimum	0.0000	0.0000	0.1000	0.1000	0.1000
Maximum	0.9956	0.9956	0.9000	0.9000	0.9000

[a]Surveyed and interpolated SU-541 municipalities- and rescaled [0.1, 0.9] scores

[b]Standard deviation

taking into account both real (80 municipalities) and interpolated values. Other combinations of those parameters ($rp_{\max}, nb_{\min}$) offer different interpolated SU and maps, but the one selected shows the closest approximated shape to real olive-tree farm geographical distribution. Table 2 summarizes basic statistics of both original and interpolated financial risks.

3.2 Local autocorrelation

The number of categories to determine relative proximity rates among SU was 10 and LISA scores were determined for $rp_{ji} \leq 5$. Relative proximity rates greater than 5 mean very large geographical distances and we cannot consider that SU with $rp_{ji} > 5$ can influence autocorrelation scores of the SU under study. As expected, the spatial distributions of individual LISA scores are very different, difficult to interpret and spatially in conflict (Fig. 2b,

c and d). Table 2 summarizes some LISA score basic statistics once they were rescaled in a range [0.1, 0.9] to avoid undesirable negative values for the objectives in MOEA.

3.3 MOEA structure and parameters

The initial population, which was designed 4 times to guarantee the geographical diversity, had 400 solutions. For each fitness function, the MOEA was run twice to determine two hot-spot sets (Sect. 2.3). The crossover rate was 0.1 (double was 0.005) and the mutation rate was 0.05. These rates were selected after several tests (1000 generations). These tests showed crossover to be more useful than mutation in achieving efficient solutions as the solution minimum distance did not change a lot. Once a SU has been selected to be mutated, both its distance-based mutation rate and fitness-based mutation rate were 0.25. Crossover rates were higher than mutation rates because the first operator is experimentally more useful than the second to maintain heterogeneity stress in the population. The maximum number of generations was established at 10000, 2.5% was the admissible MSE (20 consecutive times to stop the algorithm) and elitism was admitted in crossover and mutation from the 10th generation.

Seven objectives were finally analysed: Moran's I mean ($\bar{I}$, rescaled values, maximized) and SD_I, Geary's C mean ($\bar{C}$, rescaled values, minimized) and SD_C, Getis and Ord's G mean ($\bar{G}$, rescaled values, maximized) and SD_G and the minimum path between all SU (P, relative distance rates, minimized). All the SD objectives were minimized. Only the minimum path was constrained to $P \leq 0.5(nSU - 1)$, nSU being the number of SU in a solution which was 5. The solutions where $nSU \geq 5$ do not offer different hot-spots and the higher the nSU the greater the computer demand. In order to evaluate the fitness MF when the fuzzy evaluation of weighted objectives was selected, 11 MFs were chosen (2 extreme Z functions and 9 triangular symmetric ones). Taking into account the resulting fitness values, the number of MFs selected is enough to manage statistical diversity.

In order to identify extreme SU in the ENDSF, 30 equal intervals were selected for Q-Q Plots (Fig. 3 shows an illustrative example). More intervals do not influence the identification of the extreme SU. Only extreme SU, identified by an exponential fit, were considered to identify and locate hot-spots on a map.

4 Results

The Mann-Whitney U test, comparing original and interpolated financial risks, showed that there is no significant difference between the two means (p-value = 0.095) and, therefore, the estimation of LISA scores using both original and spatially interpolated financial risks is appropriate. Figures 2b, c and d demonstrate that the spatial projections of LISA scores show different patterns and the spatial compromises cannot be defined either visually or by using map algebra.

For all the fitness functions evaluated and runs carried out, the number of efficient solutions in the ENDSF was always over 120 (Table 3). The differences between runs and methods are not evident when analysing their basic statistics but the Mann-Whitney U test showed that these differences are noteworthy (Table 4) especially where $\bar{I}$ and $\bar{G}$ are concerned. When the efficient solutions obtained in different runs are compared, $\bar{I}$ always shows significant differences and is the most discriminant LISA score. $\bar{G}$ as well as P are relevant to obtain significant mean differences between runs when comparing the standard SPEA2 fitness function and the fuzzy evaluation of weighted objectives.

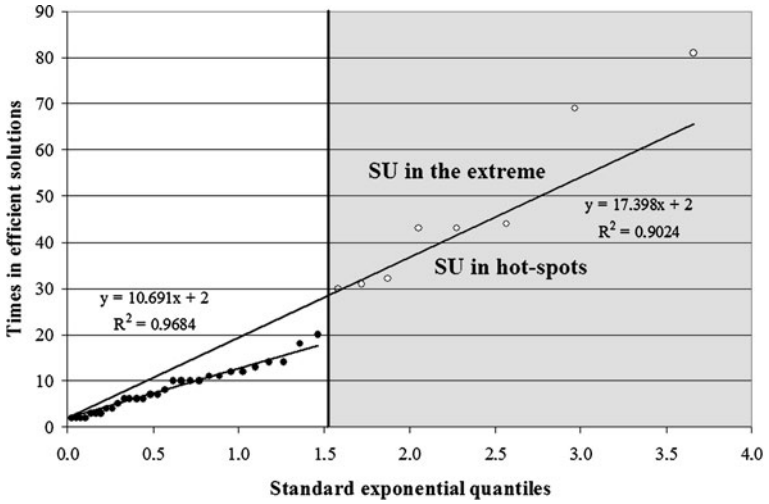

Fig. 3 Example of a Q-Q Plot based on an exponential probability distribution (30 equal intervals). Extremes identified (SPEA2, run #1, SU that appear more than one time in an efficient solution)

Table 3 Basic statistics of rescaled [0.1, 0.9] objectives—mean and (SD)—in the external non-dominated solution file. Only LISA and minimum path objectives have been included

Fitness	First MOEA run					Second MOEA run				
	$\#S$[a]	$Max\bar{I}$	$Min\bar{C}$	$Max\bar{G}$	$MinP$[b]	$\#S$[a]	$Max\bar{I}$	$Min\bar{C}$	$Max\bar{G}$	$MinP$[b]
SPEA2	128	0.761	0.127	0.778	0.119	144	0.68	0.127	0.762	0.177
		(0.066)	(0.015)	(0.039)	(0.068)		(0.146)	(0.019)	(0.043)	(0.099)
WO[c]	175	0.705	0.127	0.766	0.193	135	0.666	0.128	0.765	0.174
		(0.119)	(0.013)	(0.04)	(0.099)		(0.1)	(0.013)	(0.056)	(0.094)
FI[d]	186	0.769	0.126	0.76	0.175	158	0.738	0.126	0.770	0.19
		(0.067)	(0.013)	(0.033)	(0.098)		(0.101)	(0.014)	(0.03)	(0.093)

[a]Number of efficient solutions

[b]Minimum path between SU

[c]WO: Weighted sum of objectives weights defined at random

[d]FI: Fuzzy inference of weighted objectives

The behaviour of $\bar{C}$ is really interesting because it does not show any significant difference when fitness functions and/or runs are evaluated (Table 4). As Table 3 shows, the values of $\bar{C}$ remain practically constant in all the tests: this means that the spatial distribution of Geary's C is very constrained to specific locations. Once the MOEA identifies these places, the other LISA scores highlight specific zones. As P has been constrained, the resulting MOEA model was, on one hand, slightly less computer-time demanding and, on the other, it focused its attention on more useful solutions from a decision-making point of view.

There are not many SU in efficient solutions (Table 5) and only a few have been detected in the extreme of the Q-Q Plots. As expected, all of them show very high financial risk but they are not geographically located in the same place (Fig. 4). Two different hot-spots have

Table 4 Results of the Mann-Whitney U (p-values) test comparing LISA and minimum path objectives

		$Max\bar{I}$	$Min\bar{C}$	$Max\bar{G}$	$MinP$
SPEA2 #1	SPEA2 #2	0.000[a]	0.417	0.000[a]	0.000[a]
SPEA2 #1	WO #1	0.000[a]	0.951	0.013[b]	0.000[a]
SPEA2 #1	WO #2	0.000[a]	0.457	0.051	0.000[a]
SPEA2 #1	FI #1	0.093	0.938	0.000[a]	0.000[a]
SPEA2 #1	FI #2	0.321	0.764	0.014[b]	0.000[a]
SPEA2 #2	WO #1	0.248	0.215	0.281	0.592
SPEA2 #2	WO #2	0.005[a]	0.083	0.021[b]	0.344
SPEA2 #2	FI #1	0.000[a]	0.247	0.666	0.321
SPEA2 #2	FI #2	0.000[a]	0.626	0.044[b]	0.244
WO #1	WO #2	0.000[a]	0.406	0.15	0.116
WO #1	FI #1	0.000[a]	0.878	0.076	0.086
WO #1	FI #2	0.012[b]	0.521	0.361	0.248
WO #2	FI #1	0.000[a]	0.371	0.001[a]	0.945
WO #2	FI #2	0.000[a]	0.28	0.133	0.006[a]
FI #1	FI #2	0.007[a]	0.623	0.005[a]	0.003[a]

[a]Significant at $\alpha = 0.01$

[b]Significant at $\alpha = 0.05$

Table 5 Number of spatial units (SU)—municipalities—in hot-spots/(total SU included in efficient solutions) and basic statistics of rescaled $[0.1, 0.9]$ financial risks of hot-spot SU

	First MOEA run			Second MOEA run		
Fitness	#SU	Mean[a]	SD[a]	#SU	Mean[a]	SD[a]
SPEA2	8/(64)	0.8707	0.0280	5/(76)	0.8692	0.0218
WO	8/(101)	0.8528	0.0725	4/(100)	0.8665	0.0047
FI	6/(79)	0.8797	0.0191	5/(75)	0.8889	0.0450

[a]Rescaled financial risk

been identified in Andalusia one in the west and the other in the east. There are relevant differences in the geographical distribution of hot-spot SU when fitness function spatial projections are compared, but their locations remain almost constant. The western hot-spot is larger than the eastern one; it includes small SU that are always identified in the first run. The structure and spatial location of the second hot-spot in the east is more variable because only one SU has been detected by all the MOEA fitness functions and runs.

As expected, the structure of hot-spots is fuzzy when r fitness functions are used and n runs are carried out. After the first run, whatever the fitness function selected, the original autocorrelation values of the SU in the first-frontier hot-spots were replaced by their corresponding means calculated excluding them. In our specific problem, the SU in the first hot-spot set do not appear in the second. We cannot affirm that this phenomenon can be generalized and, depending on the original variable values, the same SU could be identified in different runs. $X_{SU_j i}$ is the number of times the SU_j is considered as a part of a hot-spot in the run i $(X_{SU_j i} \leq r)$ and each run is weighted with the inverse of the run number $w_i = 1/i$,

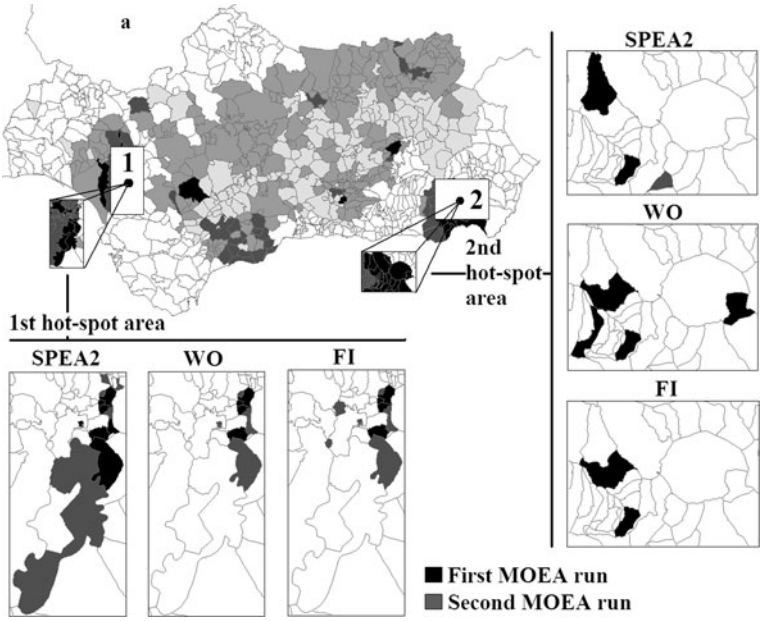

Fig. 4 Hot-spots (areas 1 and 2) obtained comparing the results of three different fitness functions and two MOEA runs. The rescaled financial risk map (**a**) is shown as a reference

the SU_j membership degree $\mu(SU_j)$ is: (i) 1 if $X_{SU_j 1} = r$; (ii) if not, then

$$\mu(SU_j) = \sum_{i=1}^{n} w_i X_{SU_j i} \Big/ r \sum_{i=1}^{n} w_i \quad \forall j \tag{4}$$

This fuzzy approach highlights the geographical zones where the phenomenon is significantly concentrated and minimizes the existence of dominant SU. In the first hot-spot four SU and in the second only one has a $\mu(SU_j) = 1$ (Fig. 4). On the other hand and considering the first hot-spot, there are two SU detected in the first run using the SPEA2 fitness function that have also been detected by the others fitness functions (the weighted sum of objectives and the fuzzy inference of weighted objectives) in the second run. In this case both have a membership degree $\mu(SU_j) = (1 + 2 \times 1/2)/[3 \times (1 + 1/2)] = 0.\widehat{4}$.

5 Conclusions

MOEA have demonstrated their utility and reliability in determining hot-spots in a complex framework like the spatial distribution of financial risk in olive-tree farms in Andalusia. LISA scores can be spatially analysed in a GIS but show different projections in spatial conflict that are difficult to interpret. To analyse econometric models that are to be considered as objectives, a MOEA can approximate the Pareto set when each efficient solution is understood as a non-repeated, close and highly autocorrelated group of SU: a potential hot-spot. Not all of the SU in efficient solutions can be considered to be part of a hot-spot, especially when their probabilities of being in an efficient solution are very low. This kind

of SU is very frequent and makes an additional analysis necessary in order to identify real SU in hot-spots. Q-Q Plots can be used to determine SU in hot-spots when an exponential probability distribution is considered to adjust the statistical distribution of extremes. This procedure assures the identification of very relevant SU but also discards others that could be relevant as well. In order to identify new hot-spots, the MOEA can be run again, but this time the means of the corresponding LISA values substitute the original LISA scores of previously identified hot-spot SU. This procedure is computer-demanding but offers very robust estimations of those SU that can be included in hot-spots.

Some LISA objectives do not give any statistical variability when results of different MOEA strategies and/or runs are compared. Very different mean values of other objectives can be obtained without substantially modifying the mean of the first ones (this is especially true when $\bar{C}$ is analysed). In the end, these objectives did not give statistical variability in the Pareto set, although they did guide the MOEA algorithm, and only when they reached a specific range of values did the other objectives vary.

In our context, it is not possible to determine if one specific MOEA fitness function is better than another because the dominance concept cannot easily be extended to evaluate the models. For example, in the first run, SPEA2 means are always better than those shown by the weighted sum of objectives (Table 3). In this case, only SD_C is a little bit lower, although this circumstance is different when the second run means are compared.

The statistical identification of SU in hot-spots is not enough; using GIS utilities, decision makers can locate them geographically. The existence of different MOEA fitness functions always provokes differences in the location of hot-spots that are related to the presence or absence of specific SU but, at least in our example, the geographical hot-spot locations remain relatively constant. This means that it is possible to achieve a compromise between different econometric methods highlighting very special zones where a specific phenomenon is highly autocorrelated. The hot-spot shapes and the number of SU in them can be improved by repeating the process "removing" previous selections statistically. Therefore, hot-spot structures can be considered fuzzy sets where each SU has a membership degree that can be obtained taking into account the number of times that it appears in a hot-spot for each run.

Acknowledgements The author acknowledges the financial subsidy provided by the Spanish Department of Research of the Ministry of Education and Science under the TIN2005-08386-C05-02 project, by the Junta de Andalucía under P05-TIC-00531 and P08-TIC-3745 projects and by the Ministry of Health Care under the PI08/90752 project. FEDER also provided additional funding. Moreover, the author would like to thank the Consejería de Agricultura y Pesca (Junta de Andalucía) for its financial and technical support.

References

Agarwal, D. K., Silander, J. A., Gelfand, A. E., Dewar, R. E., & Mickelson, J. G. (2005). Tropical deforestation in Madagascar: analysis using hierarchical, spatially explicit Bayesian regression models. *Ecological Modelling*, *185*, 105–131.

Andrienko, N., & Andrienko, G. (2006). *Exploratory analysis of spatial and temporal data: a systematic approach*. Berlin: Springer.

Anselin, L. (1995). Local indicators of spatial association: LISA. *Geographical Analysis*, *27*(2), 93–115.

Anselin, L. (2002). Under the hood: issues in the specification and interpretation of spatial regression. *Models of Agricultural Economics*, *27*, 247–267.

Asai, K. (1995). *Fuzzy systems for information processing*. Amsterdam: IOS Press.

Assunçao, R., Costa, M., Tavares, A., & Ferreira, S. (2006). Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, *25*, 723–742.

Bäck, T. (1996). *Evolutionary algorithms in theory and practice*. New York: Oxford University Press.

Barbosa, H. J., & Barreto, A. M. (2001). An interactive genetic algorithm with co-evolution of weights for multiobjective problems. In L. Spector, E. D. Goodman, A. Wu, W. Langdon, H. M. Voigt, M. Gen, S. Sen, M. Dorigo, S. Pezeshk, M. H. Garzon, & E. Burke (Eds.), *Proceedings of the genetic and evolutionary computation conference (GECCO'2001)* (pp. 203–210), San Francisco, California. San Mateo: Morgan Kaufmann.

Baddeley, A., Gregori, P., Mateu, J., Stoica, R., & Stoyan, D. (2006). *Case studies in spatial point process modeling*. Berlin: Springer.

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. (2004). *Statistics of extremes: theory and applications*. Chichester: Wiley.

Coello-Coello, C. A., & Lamont, G. B. (2004). *Applications of multi-objective evolutionary algorithms*. Singapore: Word Scientific.

Coello-Coello, C. A., Lamont, G. B., & Van Veldhuizen, D. A. (2007). *Evolutionary algorithms for solving multi-objective problems*. New York: Springer.

Das, I., & Dennis, J. (1997). A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems. *Structural Optimization*, *14*(1), 63–69.

De Pinto, A., & Nelson, G. C. (2007). Modelling deforestation and land-use change: sparse data environments. *Journal of Agricultural Economics*, *58*(3), 502–516.

Deb, K. (1999). *Non-linear goal programming using multi-objective genetic algorithms* (Technical Report CI-60/98). Dortmund: Department of Computer Science/LS11, University of Dortmund, Germany.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA–II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197.

Emir-Farinas, H., & Francis, R. L. (2005). Demand point aggregation for planar covering location models. *Annals of Operation Research*, *136*, 175–192.

Fischer, M. M. (2006). *Spatial analysis and geocomputation*. Berlin: Springer.

García-Alonso, C. R., & Pérez-Naranjo, L. (2007). Identification of financially compromised areas in the agrarian sector. In *22nd European conference on operational research book of abstracts* (Vol. 1, p. 75).

Gonçalves, G. E., & Estellita, L. P. (2002). Integrating geographical information systems and multi-criteria methods: a case study. *Annals of Operation Research*, *116*, 243–269.

Griffith, D. A. (2003). Using estimated missing spatial data with the 2-median model. *Annals of Operation Research*, *122*, 233–247.

Gutiérrez, P. A., López-Granados, F., Peña-Barragán, J. M., Jurado-Expósito, M., Gómez-Casero, M. T., & Hervás-Martínez, C. (2008). Mapping sunflower yield as affected by *Ridolfia segetum* patches and elevation by applying evolutionary product unit neural networks to remote sensed data. *Computers and Electronics in Agriculture*, *60*(2), 122–132.

Hajela, P., & Lin, C. Y. (1992). Genetic search strategies in multicriterion optimal design. *Structural Optimization*, *4*, 99–107.

Hodgson, M. J., & Jacobsen, S. K. (2009). A hierarchical location-allocation model with travel based on expected referral distances. *Annals of Operation Research*, *167*, 271–286.

Hof, J., & Bevers, M. (2000). Direct spatial optimization in natural resource management: four linear programming examples. *Annals of Operation Research*, *95*, 67–81.

Holloway, G., Lacombe, D., & LeSage, J. P. (2007). Spatial econometric issues for bio-economic and land-use modelling. *Journal of Agricultural Economics*, *58*(3), 549–588.

Horn, J., Nafpliotis, N., & Goldberg, D. E. (1994). A niched Pareto genetic algorithm for multiobjective optimization. In *Proceedings of the first IEEE conference on evolutionary computation, IEEE World congress on computational intelligence* (Vol. 1, pp. 82–87). Piscataway: IEEE Service Center.

Ishbuchi, H., & Murata, T. (1996). Multi-objective local search algorithm and its applications to flowshop scheduling. *IEEE Transactions on Systems, Man, and Cybernetics*, *28*(3), 392–403.

Ishibuchi, H., & Murata, T. (1998). Multi-objective genetic local search algorithm and its application to flowshop scheduling. *IEEE Transactions on Systems, Man, and Cybernetics*, *28*(3), 392–403.

Junta de Andalucía (2004). Anuario de estadísticas agrarias y pesqueras año 2004. Consejería de Agricultura, Pesca y Alimentación. http://www.juntadeandalucia.es/agriculturaypesca/portal/www/portal/com/bin/portal/DGPAgraria/Estadisticas/estadisticasagrarias/anuario2004.pdf. Accessed 19 November 2007.

Kalyanmoy, D. (2004). *Multi-objective optimization using evolutionary algorithms*. Chichester: Wiley.

Kursawe, F. (1991). A variant of evolution strategies for vector optimization. In H. P. Schwefel & R. Männer (Eds.), *Lecture notes in computer science: Vol. 496. Parallel problem solving from nature. 1st workshop, PPSN I* (pp. 193–197), Dortmund, Germany. Berlin: Springer.

Laumanns, M., Thiele, L., & Zitzler, E. (2006). An efficient, adaptive parameter variation scheme for metaheuristics based on the epsilon-constraint method. *European Journal of Operational Research*, *169*, 932–942.

Lee, M. A., & Esbensen, H. (1997). Fuzzy/multiobjective genetic systems for intelligent systems design tools and components. In W. Pedrycz (Ed.), *Fuzzy evolutionary computation*. Boston: Kluwer Academic.

Loughlin, D. H., & Ranjithan, S. (1997). The neighborhood constraint method: a genetic algorithm-based multiobjective optimization technique. In T. Bäck (Ed.), *Proceedings of the seventh international conference on genetic algorithms* (pp. 666–673). San Mateo: Morgan Kaufmann.

Michalewicz, Z., & Schoenauer, M. (1996). Evolutionary algorithms for constrained parameter optimization problems. *Evolutionary Computation*, *4*(1), 1–32.

Moreno, B., García-Alonso, C. R., Negrín-Hernández, M. A., Torres-González, F., & Salvador-Carulla, L. (2008). Spatial analysis to identify hotspots of prevalence of schizophrenia. *Social Psychiatry and Psychiatric Epidemiology*. doi:10.1007//s00127-008-0368-3.

Noyan, N. (2010). Alternate risk measures for emergency medical service system design. *Annals of Operation Research*. doi:10.1007/s10479-010-0787-x.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, *27*, 286–306.

Pérez-Naranjo, L. M., & García-Alonso, C. R. (2005). Spatial income distribution of horticultural farms in Andalusia. *Cuadernos Geográficos*, *37*, 41–58.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Smith, A. E., & Coit, D. W. (1997). Constrain handling techniques—penalty functions. In T. Bäck, D. B. Fogel, & Z. Michalewicz (Eds.), *Handbook of evolutionary computation*. London: Oxford University Press.

Srigiriraju, K. C. (2000). *Noninferior surface tracing evolutionary algorithm (NSTEA) for multi objective optimization*. Master's thesis, North Carolina State University, Raleigh, North Carolina.

Staal, S. J., Baltenweck, L., Waithaka, M., de Wolff, T., & Njoroge, L. (2002). Location and uptake: integrated household and GIS analysis of technology adoption and land use with application to smallholder dairy farms in Kenya. *Agricultural Economics*, *27*, 295–315.

Tavares-Pereira, F., Rui Figueira, J., Mousseau, V., & Roy, B. (2007). Multiple criteria districting problems: the public transportation network pricing system of the Paris region. *Annals of Operation Research*, *154*, 69–92.

Wang, G., & Terpenny, J. P. (2005). Interactive preference incorporation in evolutionary engineering design. In Y. Jin (Ed.), *Knowledge incorporation in evolutionary computation*. Berlin: Springer.

Wilson, P. B., & MacLeod, M. D. (1993). Low implementation cost IIR digital filter design using genetic algorithms. In *IEE/IEEE workshop on natural algorithms in signal processing* (pp. 4/1–4/8). Chelmsford, UK.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

Youssef, H. A., Kinsella, A., & Waddington, J. L. (1991). Evidence for geographical variations in the prevalence of schizophrenia in rural Ireland. *Archives of General Psychiatry*, *48*, 254–258.

Zitzler, E., & Thiele, L. (1999). Multiobjective evolutionary algorithms: a comparative case study and strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, *3*(4), 257–271.

Zitzler, E., Laumanns, M., & Thiele, L. (2001). SPEA2: improving the strength of Pareto evolutionary algorithm. In K. Giannakoglou, D. Tsahalis, J. Periaux, P. Papailou, & T. Fogarty (Eds.), *Evolutionary methods for design, optimization and control with applications to industrial problems (EUROGEN 2001)*, Athens.

Zydallis, J. B., Veldhuizen, D. A. V., & Lamont, G. B. (2001). Statistical comparison of multiobjective evolutionary algorithms including the MOMGA–II. In E. Zitzler, K. Deb, L. Thiele, C. A. C. Coello, & D. Corne (Eds.), *Lecture notes in computer science: Vol. 1993. First international conference on evolutionary multi-criterion optimization* (pp. 226–240). Berlin: Springer.