A Permutation-Randomization Approach to Test the Spatial Distribution of Plant Diseases

Published version:

DOI:10.1094/PHYTO-05-15-0112-R

[Lione G., Gonthier P., 2016. Phytopathology, 106, pp. 19–28,
DOI: 10.1094/PHYTO-05-15-0112-R]

A permutation-randomization approach to test the spatial distribution of plant diseases

G. Lione, and P. Gonthier

Department of Agricultural, Forest and Food Sciences, University of Torino, Largo P. Braccini 2, 10095 Grugliasco, Italy.

Corresponding author: P. Gonthier; E-mail address: paolo.gonthier@unito.it

ABSTRACT

G. Lione, and P. Gonthier, 2015. A permutation-randomization approach to test the spatial distribution of plant diseases. Phytopathology xx:xxxx-xxxx.

The analysis of the spatial distribution of plant diseases requires the availability of trustworthy geostatistical methods. The MDT (Mean Distance Tests) are here proposed as a series of permutation and randomization tests to assess the spatial distribution of plant diseases when the variable of phytopathological interest is categorical. A user-friendly software to perform the tests is provided. Estimates of power and type I error, obtained with Monte Carlo simulations, showed the reliability of the MDT (power>0.80; type I error<0.05). A biological validation on the spatial distribution of spores of two fungal pathogens causing root rot on conifers was successfully performed by verifying the consistency between the MDT responses and previously published data. An application of the MDT was carried out to analyze the relation between the plantation density and the distribution of the infection of *Gnomoniopsis castanea*, an emerging fungal pathogen causing nut rot on sweet chestnut. Trees carrying nuts infected by the pathogen were randomly distributed in areas with different plantation densities, suggesting that the distribution of *G.*

castanea was not related to the plantation density. The MDT could be used to analyze the spatial distribution of plant diseases both in agricultural and natural ecosystems.

Additional keywords: geostatistics, *Gnomoniopsis castanea*, Mean Distance Tests, permutation, randomization, resampling, spatial pattern.

INTRODUCTION

Analyzing the spatial pattern of plant diseases may be pivotal to elucidate the ecology, the epidemiology and the infection biology of pathogens as well as the mechanisms underlying host-pathogen interactions and the spread of epidemics (Nelson et al. 1999). A large body of literature deals with the application of Geographic Information Systems (GIS) in conjunction with statistical and geostatistical methods to investigate peculiar traits of plants diseases, to test biologically relevant hypotheses and to build predictive and/or explicative models (Nelson et al. 1999). Examples of GIS and geostatistical applications can be found in both agriculture and forestry on a broad range of diseases, hosts and pathogens, including viruses, bacteria and fungi. For instance, GIS and geostatistical analyses were used to relate the presence of tomato virus vectors to the spatial pattern of the symptoms in tomato (*Solanum lycopersicum* L.) crops (Nelson et al. 1999). Analogous analyses were performed to test the association between genetic variations in cotton leaf curl viruses and the disease severity in *Gossypium* spp. fields (Nelson et al. 1999) and to investigate the dispersion mechanisms of the plum pox potyvirus in orchards of *Prunus armeniaca* L. and *P. persica* (L.) Batsch (Gottwald et al. 1995). Similar approaches were carried out to elucidate the role of pedoclimatic factors on the incidence of the bacterial blight caused by *Xanthomonas arboricola* pv. *corylina* on *Corylus avellana* L. (Lamichhane et al. 2013). GIS and geostatistics were also used to explore the spatial distribution of genotypes of *Phytophthora infestans* (Mont.) de Bary in orchards of *S. lycopersicum* and *Solanum tuberosum* L. affected by late blight disease (Jaime-

Garcia et al. 2000) and of *P. nicotianae* B. de Haan var. *parasitica* (Dast.) Waterh. in crops of *Ananas comosus* (L.) Merr. (Chellemi et al. 1988). A GIS and geostatistical-based technique was used to model the spatio-temporal dynamics of the leaf spot associated with *Ramularia areola* G. F. Atk. in *Gossypium* spp. crops (Pizzato et al. 2014) and to test the relation between climatic factors and the incidence of the nut rot caused by *Gnomoniopsis castanea* Tamietti in orchards of *Castanea sativa* Miller (Lione et al. 2014). GIS and geostatistics were also applied to the study of the ecological association between the alien forest pathogen *Heterobasidion irregulare* Garbel. & Otrosina and the habitats of its invasion area in Europe (Gonthier et al. 2012), as well as to define adequate management prescriptions to thwart the invasion (Gonthier et al. 2014).

As shown in this overview, regardless of the spatial scale of the study and of the pathosystem under investigation, many experimental designs in plant pathology are characterized by a recurring pattern. Within this pattern, points (e.g. individual plants, sampling sites or spore trapping devices) are defined by spatial coordinates and by a variable of phytopathological relevance. This variable can be either quantitative (e.g. disease incidence, disease severity, amount of inoculum) or categorical (e.g. infected/healthy plant, plant showing heavy/moderate/mild symptoms, infested/not infested site). The analysis of the spatial distribution of points and of the associated variable relies on different conceptual and computational approaches.

Several methods are available to assess whether the spatial distribution of points is clustered, random or dispersed, including the Nearest Neighbor Index (NNI), the Ripley’s K function and the Nearest Neighbor Hierarchical Clustering (NNHC), whose significance is generally estimated with Monte Carlo (MC) simulations (Mitchell 2009). The rationale of MC simulations lies in the comparison between the observed points location and the location of a large number of points samples drawn from a predefined data generating process (DGP) known as point process (Crawley 2013; Carsey and Harden 2014). The choice of the appropriate point process depends upon the null hypothesis being tested (de Smith et al. 2007).

The spatial distribution of the quantitative variable associated with points is generally assessed through spatial autocorrelation analyses involving the Mantel test, the estimation of variograms and the calculation of autocorrelation indexes such as the Geary's c, the Moran's I and the Getis-Ord general G-statistic at global or local scale (Mantel 1967; Mitchell 2009; Webster and Oliver 2001). To account for the stochastic uncertainty related to these methods, asymptotic theory and heuristic procedures are available (Goslee and Urban 2007; Marchant and Lark 2004; Mitchell 2009). While the above cited techniques are routinely applied and embedded in some major GIS and statistical software (Mitchell 2009), the spatial distribution of a categorical variable associated with points is still a topic of active research and ongoing development. In the last decades plant pathologists have proposed and validated some conceptual and technical solutions to this issue. For instance, the software package 2DCLASS was designed to perform the Gray's analysis aimed at detecting the spatial pattern of plant diseases (Gray et al. 1986; Nelson et al. 1992). 2DCLASS was further improved by the STCLASS package (Nelson 1995) and by a MC-based approach to investigate the spatiotemporal pattern of the spread of epidemics (Thébaud et al. 2005). A correlation-based technique was also proposed to detect the spatial distribution of discrete data through the 2DCORR package (Ferrandino 1997). More recently, an extension of local measures of spatial association was suggested to deal with the same kind of data (Boots 2003). The above cited solutions were designed to analyze binomial categorical data (e.g. infected/healthy plant) in lattices, where points were approximated to cells in a regular grid, including missing points (e.g. missing plants). While this approximation is suitable to model many field conditions where plants are located in the space according to a predefined geometric pattern, like in nurseries, in orchards and in regular plantations, no application to forestry, to irregular plantations and to natural seedlings regeneration has been reported so far. Despite transiogram analyses were proposed to overcome the constraints related to the plants plantation scheme, the discrepancy between experimental transiograms and idealized ones can occur, affecting the interpretation of the results (Weidong 2006).

The goal of this study was to develop and validate a permutation and randomization-based approach, hereafter called Mean Distance Tests (MDT), to assess the spatial pattern of a plant disease when this is defined as a categorical variable. The MDT algorithms were embedded in a user-friendly application for personal computer.

MATERIALS AND METHODS

Overview and software design. Let $T = \{t_1, t_2, ..., t_n\}$ be a finite set of n points with known x and y coordinates in a Cartesian plane and let $I \subset T$ be a subset of T including m $(2 \le m \le n-2)$ points. For instance, the points in the set T could be plants and the points in the subset I could be the plants infected by some pathogens. In other terms, the m points in the subset I are those points of the set T which a level γ (i.e. “infected”) of a categorical variable Γ (e.g. “health status”) has been assigned to. Let $\bar{d}$ be an overall index of the distances that separate m points in a plane, calculated as the mean of the values stored in the $m \times m$ triangular Euclidean distance matrix of the points. Let $\bar{d}_0$ be the observed value of $\bar{d}$, which is calculated for the m points included in the subset I. Finally, let be $\binom{n}{m}$ a binomial coefficient, representing the number of possible arrangements of m elements drawn from a set of n elements. Within the permutation tests framework, the probability mass function (PMF) of $\bar{d}$ is obtained by calculating $\bar{d}$ for each i^{th} combination $\left[1 \le i \le \binom{n}{m}\right]$ through which m points of the set T can be randomly assigned to the subset I (Carsey and Harden 2014). Instead, within the randomization tests framework, the PMF is estimated by calculating $\bar{d}$ on a random sample without replacement of B combinations $\left[1 < B < \binom{n}{m}\right]$ (Carsey and Harden 2014). The main core of this work is to determine from the PMF, with a predefined significance level cut-off α, whether $\bar{d}_0$ is either significantly lower (i.e. located towards the left tail) or higher (i.e. located towards the right tail) than expected under the random assignment of γ (i.e. random definition of the subset I within the set T). The first case indicates a clustered spatial pattern of the level γ, while the second occurs in a dispersed spatial pattern of the same level. This is equivalent to test if the infected plants are nearer or further apart than expected according to a random distribution of the infected plants within the sampled plants. To deal with this issue the Mean Distance Tests (MDT) approach is proposed here.

MDT are based on the assumption that the x and y coordinates of points in the set T are fixed and that only the assignment of the level γ is a stochastic process. The MDT consist of 3 permutation tests (Mean Distance Permutation Tests - MDPT) and 3 randomization tests (Mean Distance Randomization Tests - MDRT). Both permutation and randomization tests are divided according to the tails of the PMF they refer to (Hartwig 2013). MDPT2T is the two-tailed (2T) permutation test, MDPTLT the left-tailed (LT) and MDPTRT the right-tailed (RT), respectively. Similarly, the MDRT are designed in the two-tailed version (MDRT2T), in the left-tailed (MDRTLT) and in the right-tailed (MDRTRT) ones (Table 1). Once the above described steps to obtain the PMF and to calculate $\bar{d}_0$ are performed, the mean value $\bar{D}$ of the PMF is calculated, the exact p-value (p_e) is determined for MDPT and the randomization p-value (p_r) is determined for MDRT as reported in Carsey and Harden (2014) and Ernst (2004). The adequacy of the number B selected to perform the MDRT is assessed by calculating the lower (L_{pr}) and upper (U_{pr}) bounds of the confidence interval for p_r at user-defined level λ (e.g. 0.95). The confidence interval is calculated from the binomial distribution as described in Ernst (2004). Whenever the condition $L_{pr} \le \alpha \le U_{pr}$ is verified, p_r is deemed to be ambiguous and B is increased until the sampling adequacy is achieved and, thus, ambiguity is solved (Ernst 2004).

The algorithms performing the MDT were compiled and run in R 3.1.2 environment (R Core Team, Vienna, Austria) and subsequently embedded in a software for personal computer designed with Shiny, a hybrid R-HTML environment for personal computer (Beeley 2013).

Monte Carlo estimates of MDPT power and type I error. MC simulations were performed to assess the power and the type I error of MDPT2T, MDPTLT and MDPTRT. According to the null hypothesis of each test (Table 1), three DGPs were designed. Every DGP consisted in a point process realized both in a squared 4 × 4 units window and in a 6 × 6 one. The point processes included n=15 points for the set T and from m=2 to m=13 points for the subset I. The origin of the Cartesian system was located in the windows centre and the points coordinates were expressed in polar form (R, θ). The first DGP (point process 1 - PP1) was designed to simulate a random spatial distribution of γ. At each MC simulation, the set T was generated by sampling for n times R from a uniform distribution (Carsey and Harden 2014) bounded between 0 and half the window edge and θ from a uniform distribution bounded between 0 and 2π radians. A random number generator was used to define the subset I by drawing m out of n points without replacement, with the extraction probability set constant for each point (Carsey and Harden 2014). The level γ was assigned to the sampled m points. The second DGP (PP2) was planned to simulate a clustered spatial distribution of γ. The level γ was assigned to m points whose R was sampled from a beta distribution with shape parameters a=0.5 and b=10 (Crawley 2013) and whose θ was generated from the same uniform distribution described for PP1. The remaining points were drawn in the same way but inverting the a and b shape parameters. In the last DGP (PP3) a dispersed spatial distribution of γ was simulated. PP3 was set as described for PP2 with the exception of the shape parameters of the beta distribution, which were inverted.

To gather the estimates of permutation tests power and type I error, two blocks of MC simulations (hereafter blocks), each one consisting in $1 \cdot 10^4$ simulations, were performed for both windows, for every m value and for any MDPT, resulting in a total of $1.44 \cdot 10^6$ simulations. For each block either a single DGP or a couple of DGPs selected among PP1, PP2 and PP3 was run.

The number of simulations based on PP1, PP2 or PP3 within a single block varied depending on the MDTP (Table 2). For every simulation within the block the same permutation test was performed on the γ level with the α value set to 0.05. As proposed by Thébaud et al. (2005), the proportion of simulations resulting in the rejection of a false null hypothesis was used as an estimate of power. Similarly, the estimate of type I error was calculated as the proportion of simulations within a single block in which MDPT rejected the null hypothesis when it was true. The estimates of power and type I error were averaged to be compared among tests and windows size. The above estimates were also correlated with the Spearman ρ correlation coefficient to m [i.e. testing $\rho(m)$] and to $\binom{n}{m}$ [i.e. testing $\rho\binom{n}{m}$], with a p-value cut-off set to 0.05.

Biological validation. The MDRT were validated on data gathered from Gonthier et al. (2012). In this study, 44 sampling points equipped with spores trapping devices were located within a 3030 ha forest in the Circeo National Park, in central Italy. Spore trapping devices allowed to determine the spores deposition rate (DR), expressed as the number of viable spores per squared meter per hour (spores$\cdot$m$^{-2}\cdot$h^{-1}), of two fungal pathogens causing root rot on conifers. The first pathogen, *Heterobasidion annosum* (Fr.) Bref., is native in the area, while the second one, *H. irregulare,* is an alien invasive species. Geostatistical analyses of spatial autocorrelation performed on the DR showed that *H. irregulare* was ubiquitous and distributed in the area according to a random spatial pattern, while *H. annosum* showed significant clustering around patches of conifers.

To validate the MDRT, the set T was defined including all n=44 sampling points. Two categorical variables Γ_1 (i.e. “presence of *H. annosum* spores”) and Γ_2 (i.e. “presence of *H. irregulare* spores”) were defined. For Γ_1 the γ_1 level (i.e. “*H. annosum* spores are present”) was assigned to the m_1 sampling points with *H. annosum* DR>0, which were included in the subset I_1. Similarly, the γ_2 level (i.e. “*H. irregulare* spores are present”) was assigned to the m_2 sampling points with *H. irregulare* DR>0 to define the subset I_2. MDRT2T, MDRTLT and MDRTRT with α=0.05, B=10^4 and λ=0.95 were performed on both γ_1 and γ_2 levels.

Application to a case study. An application of the MDT to a case study was carried out to test the relation between the plantation density and the incidence of *Gnomoniopsis castanea*, an emerging fungal pathogen causing the nut rot of chestnut (Visentin et al. 2012). During October 2013, the coordinates of 203 sweet chestnuts (*C. sativa*) were recorded in UTM WGS84 zone 32 N system (m) with a GPS device (Magellan Mobile Mapper 6, Magellan Navigation Inc., Santa Clara, CA, USA). The trees grew in the sweet chestnut orchard "Vivaio Gambarello", set in the north-west of Italy (E 394,925; N 4,906,885). A NNHC analysis (Mitchell 2009) was performed on CrimeStat 3.3. (Ned Levine & Associates, Houston, TX, USA) with $2\cdot10^3$ iterations and significance level cut-off set to 0.05. The two clusters of sweet chestnuts including the largest number of trees (areas C1 and C2, see results) were selected and two not clustering groups (areas NC1 and NC2) with the same number of sweet chestnuts were randomly chosen. The mean value of the triangular Euclidean distance matrix among all the sweet chestnuts was calculated for areas C1, C2, NC1 and NC2. Up to 40 nuts per tree were collected from the crown of each sweet chestnut in the above mentioned areas. Fragments of the nuts kernel were plated in Petri dishes on Malt Extract Agar (MEA) to assess the presence/absence of *G. castanea* in the fruit tissues at the tree level. Isolations and fungal identification were performed as described by Lione et al. (2014). The incidence of *G. castanea* was calculated as the ratio, in percent, between the m_{C1}, m_{C2}, m_{NC1} and m_{NC2} trees carrying at least one infected nut (i.e. subsets I_{C1}, I_{C2}, I_{NC1} and I_{NC2} of areas C1, C2, NC1 and NC2) and the n_{C1}, n_{C2}, n_{NC1} and n_{NC2} trees growing in each area (i.e. sets T_{C1}, T_{C2}, T_{NC1} and T_{NC2}). The categorical variable Γ (i.e. "presence of *G. castanea* in at least one nut") was defined and the level γ (i.e. "*G. castanea* is present in at least one nut") was assigned to the m_{C1}, m_{C2}, m_{NC1} and m_{NC2} trees. The incidence of the pathogen was compared among the four above mentioned areas with a χ^2 test performed with a significance cut-off of 0.05. For each area $\bar{d}_0$ and $\binom{n}{m}$ were calculated. MDRT2T, MDRTLT and

MDRTRT with α=0.05 and MDRT2T, MDRTLT and MDRTRT with α=0.05, B=10^2, B=$5\cdot10^2$ and λ=0.95 were performed on the γ level for every area.

RESULTS

Software design. MDT algorithms are provided as scripts to run in R environment (Supplementary file 1). The algorithms have also been embedded in the MDT software, a "point-and-click" graphic user interface (GUI) running on the internet browser. The user is supposed to provide the input data as a spreadsheet .csv file with as many rows as the points in the set *T*, one column for each spatial coordinate, one column for the Γ variable. Cells included in this last column indicate for all points the assigned levels of Γ. The other inputs required (Table 1) should be specified directly in the GUI. The MDT software, its user manual and the installation instructions are freely available from the *e*-Xtras (Supplementary file 2).

Monte Carlo estimates of MDPT power and type I error. On average the estimates of power of MDPT ranged from 0.8884 to 0.9917, while the estimates of type I error were comprised between 0.0247 and 0.0496 depending on the test. The maximum average power was attained by MDPTLT, followed by MDPT2T and MDPTRT. The minimum values of type I error were observed in MDPTLT and MDPTRT, followed by MDPT2T. Within the same test, the window size affected the average values of the power and of the type I error estimates resulting in a maximum absolute difference of ±0.001. Significant correlations [$\rho(m)$=0.6504; P=0.0220] were detected between the power estimates and m in MDPTLT, regardless of the window size. Significant values of $\rho\binom{n}{m}$ were observed in the correlation tests between the power estimates and $\binom{n}{m}$ in MDPT2T and MDPTRT for both windows sizes [$\rho\binom{n}{m}$>0.8600; P<0.05]. No significant correlations

(P>0.05) were observed between the estimates of type I error and either m or $\binom{n}{m}$, with the exception of MDPTRT in the 6 × 6 units window (Table 3).

Biological validation. For the variable Γ_1, the level γ_1 was assigned to $m_1=16$ sampling points that fulfilled the condition *H. annosum* DR>0, defining the subset I_1 (Fig. 1A). For γ_1, the value of $\bar{d}_0$ attained 2767 m, while $\bar{D}$ was 3449 m in MDRT2T and 3443 m in both MDRTLT and MDRTRT. Based on MDRT2T, sampling points where spores of *H. annosum* had been detected were not randomly distributed within the sampling points ($p_r=0.0122$, $L_{pr}=0.0117$, $U_{pr}=0.0158$). MDRTLT indicated a clustered spatial pattern of the points with *H. annosum* DR>0 within the sampling points ($p_r=0.0092$, $L_{pr}=0.0065$, $U_{pr}=0.0113$). Finally, MDRTRT was not significant, showing a not dispersed spatial distribution of the points with *H. annosum* DR>0 within the sampling points ($p_r=0.9892$, $L_{pr}=0.9875$, $U_{pr}=0.9918$). The subset I_2 was defined by assigning the level γ_2 of the variable Γ_2 to the $m_2=29$ points that satisfied the condition *H. irregulare* DR>0 (Fig. 1B). In this case, $\bar{d}_0$ attained a value of 3281 m, while $\bar{D}$ ranged from 3445 m in MDRTLT to 3446 m in MDRT2T and MDRTRT. MDRT2T output indicated that sampling points where spores of *H. irregulare* had been identified were randomly distributed within the sampling points ($p_r=0.2554$, $L_{pr}=0.2422$, $U_{pr}=0.2699$). According to MDRTLT, points with *H. irregulare* DR>0 were not clustered within the sampling points ($p_r=0.1278$, $L_{pr}=0.1272$, $U_{pr}=0.1402$), while the MDRTRT showed a not dispersed spatial pattern for the same points ($p_r=0.8739$, $L_{pr}=0.8636$, $U_{pr}=0.8781$). In all MDRT performed the condition $L_{pr} \le \alpha \le U_{pr}$ was not verified for $B=10^4$.

Application to the case study. The NNHC showed the presence of 24 first order clusters, comprising two to five trees, and two second order clusters (areas C1 and C2), composed by four and five first order clusters with a total of $n_{C1}=14$ and $n_{C2}=17$ sweet chestnuts, respectively (P<0.05) (Fig. 2A and 2B). The same number of trees was used to define the areas NC1 ($n_{NC1}=14$) and NC2 ($n_{NC2}=17$) (Fig. 2C and 2D). The mean value of the triangular Euclidean distance matrix among all trees attained 12.8 m in C1, 9.9 m in C2, 13.1 m in NC1 and 26.3 m in NC2. The level γ was assigned to the m_{C1}=10, m_{C2}=9, m_{NC1}=8 and m_{NC2}=11 sweet chestnuts carrying at least one nut infected by *G. castanea* (Fig. 2). The incidence of *G. castanea* was 71.4% in C1, 52.9% in C2, 57.1% in NC1 and 64.7% in NC2. The χ^2 test indicated no significant differences among the incidence level of the four areas (P=0.7312). The $\bar{d}_0$ distance ranged from 18.8 m to 32.7 m, with the lowest values observed in C1 and C2, while $\binom{n}{m}$ was comprised between 1,001 and 24,310, depending on the area. The MDT performed were never significant (p_e>0.05; p_r>0.05), regardless of the area, indicating a random (2T), not clustered (LT) and not dispersed (RT) spatial distribution of sweet chestnuts infected by *G. castanea* within the sampled trees. The *B* values were adequate to perform the MDRT since the condition $L_{pr} \le \alpha \le U_{pr}$ was not verified, with the exception of the MDRTLT carried out in NC1 for B=10^2. Increasing *B* values reduced the width of the interval [L_{pr}, U_{pr}] for every MDRT in all areas (Table 4).

DISCUSSION

The analysis of the spatial pattern of plant diseases is a pivotal issue in plant pathology since it is aimed at gathering relevant information about biological, epidemiological and ecological aspects of pathogens. In this regard, during the last decades, an increasing interest has been addressed by plant pathologists to the development and the use of statistical and geostatistical methods. It is worth noting that the majority of these methods was mainly designed to analyze specific kinds of variables in a limited range of field conditions. A large body of literature dealt with the spatial distribution of relevant phytopathological measures on the continuous or ordinal scale, while few studies were focused on the spatial pattern of categorical variables. Moreover, many researches carried out on categorical variables proposed geostatistical methods aimed at analyzing diseases in lattices and in regular plantations. The application of such methods often requires the user to own a solid background in mathematics, advanced statistics and information technology, since the algorithms performing the tests are rarely wrapped into a user-friendly “point-and-click” interface. These aspects may thwart the diffusion of some statistical and geostatistical tests in phytopathology, despite they were designed explicitly to analyze plant diseases. Within this framework, the main goal of our study was to propose the MDT as a series of geostatistical tests to assess the spatial pattern of plant diseases when the variable of phytopathological interest is categorical and to provide the user with an intuitive “point-and-click” software to perform the tests.

It is worth noting that the MDT assumptions are not constrained by the spatial pattern of the points in the set *T*, thus the MDT are virtually suitable to be applied in a wide range of situations, encompassing agricultural, forest and natural ecosystems. Unlike other geostatistical tests, the MDT do not require a grid-based approximation to represent the points location, hence they can be performed on the actual vector features of the points (e.g. shape files in a GIS environment).

The MDT are based on a permutation and randomization approach, in the acceptation proposed by Carsey and Harden (2014), and consequently they are included in the broader category of non parametric techniques known as resampling methods. These methods can be profitably employed when the stochastic process underlying the phenomenon under investigation may be assumed to be well mimicked by the resampling process (Carsey and Harden 2014). This may be often the case in plant pathology. For instance, a researcher may be interested in the investigation of the spatial distribution of plants infected by some pathogens within a regular plantation. In such a situation, the location of plants is the result of a predetermined design, while the occurrence of the pathogen may be realistically assumed as a stochastic event, which could have resulted in a different outcome depending on the random factors influencing the disease (e.g. environmental variables, inoculum pressure). In natural and semi-natural ecosystems a certain level of stochasticity is intrinsic in the distribution of plants, yet it may often be considered negligible in relation to the stochasticity involved in the epidemiological processes. Moreover, a plant pathologist is generally more interested in the dynamics of the disease rather than in the dynamics underlying the actual distribution of plants within the study area. For the above cited reasons, the MDT permute (i.e.

MDPT) or randomize (i.e. MDRT) the location of the points included in the subset I, while keeping constant the coordinates of the points in the set T. This approach equals to permute or randomize the assignment of the level γ of the categorical variable Γ to m out of n points, where m and n are the points included in the subset I and in the set T, respectively. In any case, it is up to the researcher ascertaining whether the above assumptions about the stochasticity of the phytopathological process under investigation hold reasonably true according to the experimental pattern and the goals of the study.

The algorithms proposed for the MDT are largely based on the estimation of the PMF of the distance parameter $\bar{d}$ through either permutation or randomization. Both permutation and randomization are currently considered robust and flexible standards for the assessment of the PMF of parameters lacking a solid distributional theory (Carsey and Harden 2014; Ernst 2004; Peres-Neto and Olden 2001). Whenever possible, the permutation approach should be preferred, since the randomization leads to an estimate of the permutation results, implying a higher degree of uncertainty in the response. However, permutation methods may pose heavy computational issues in terms of time consumption and technical feasibility (Ernst 2004). Combinatorics shows that, even for moderate sample sizes, the amount of data generated during a permutation test may be extremely large, requiring an excessively long time to be processed, or even exceeding the available computational power of the computer. Thus, the limits of the computer performances may impose the switch from the permutation to the randomization approach (Carsey and Harden 2014). This switch implies a cost in terms of uncertainty, that in the case of the MDRT affects the value of p_r. To deal with this issue, the calculation of confidence intervals for p_r were embedded in the MDRT algorithms as indicated by Ernst (2004). It is worth noting that the theory of resampling methods suggests that a higher accuracy in the results of randomization may be acquired by increasing the number of combinations randomly selected to perform the test (Carsey and Harden 2014; Ernst 2004). This is remarkably relevant when the randomization p-value tends to approach α, the cut-off level dividing the regions of acceptance/rejection of the null hypothesis under the estimated PMF.

In fact, if the confidence interval of the randomization p-value includes α, there is no possibility of discriminating between the two regions. As shown for *G. castanea* in this study, the ambiguity in the application of the MDRTLT to the area NC1 was solved by using a 5-fold larger value of B, that excluded the value α from the 95% confidence interval of p_r. Besides, in the same case study, the reduction of the 95% confidence interval width of p_r, as well as the trend to the convergence of the randomization results to the permutation ones could be observed empirically, in agreement with the above mentioned theory of resampling methods.

Both MDPT and MDRT were designed in the two-tailed, left-tailed and right-tailed versions. Since the points included in the subset I can be mapped on a GIS and can be visually differentiated from the rest of the points of the set T, the researcher may be induced to perform a one-tailed, rather than a two-tailed test, on the basis of the spatial pattern qualitatively observed on the map. The preference accorded to the one-tailed tests may also derive from some biologically relevant information. For instance, depending on the epidemiology and infection biology of the pathogen, the researcher could be interested in investigating either clustering or dispersion rather than randomness of the infected plants within the set of sampled plants. Separate algorithms were provided depending on the tails of the PMF, because the extension of the asymptotic approach to switch from the one-tailed p-value to the two-tailed one is not recommended (Hartwig 2013).

The null hypothesis of each test was formulated according to the general principles underlying the permutation and randomization approach (Carsey and Harden 2014; Hartwig 2013) using the statistic $\bar{d}$ as overall index of the distances that separate a set of points in a plane. The definition of $\bar{d}$ is consistent with the assumptions about the spatial differences among clustered, randomized and dispersed point patterns (Crawley 2013; Mitchell 2009) and it is included in standard statistical methods dealing with clustering problems (Aldenderfer and Blashfield 1987). Accordingly, the case study of *G. castanea* showed that the values achieved by $\bar{d}$ for all trees growing in each clustering areas were lower than the values observed in non clustering areas, despite the NNHC performed for clusters identification was based on another distance index

(Mitchell 2009). It is worth noting that the statistic $\bar{d}$ is only one among the distance measures that could have been calculated as overall index of the distances that separate a set of points in a plane, yet the comparison among different distance indexes was not a goal of this study.

The MDT do not include *ad hoc* procedures to account for scale dependency of the spatial pattern of the points in the subset *I* within the set *T*. On one side, the scale dependency should not be an issue, since the scale is non included in the definition of $\bar{d}$ and it is consequently determined by the spatial extension covered by the points of the set *T*. However, since the definition of *T* is arbitrary, the MDT approach could be applied at both global and local scale (Mitchell 2009). In the latter case, the MDT could be performed on partitions of the original set *T* including contiguous points, yet it is worth noting that the disagreement between outputs obtained from global and local applications cannot be excluded, since it was reported as a common feature in the framework of geostatistical tests (Mitchell 2009), despite it was not tested in this study.

The assessment of power and type I error of permutation tests requires an heuristic approach based on MC simulations (Peres-Neto and Olden 2001; Thébaud et al. 2005). The average and the single values obtained for power and type I error estimates of MDPT were in agreement with those reported for analogous geostatistical tests by Thébaud et al. (2005). On average the power of both two-tailed and one-tailed tests was larger than 0.80, while the type I error was lower than 0.05, as generally recommended to ensure the trustworthiness of statistical tests (Crawley 2013). The number of simulations performed within each block and the number of blocks were deemed to be largely sufficient to provide reliable estimates of the power and the type I error, in agreement with previously reported data (Carsey and Harden 2014; Ernst 2004; Thébaud et al. 2005). The window sizes seemed not to be influential on the estimates of the power and of the type I error, as demonstrated by the small differences detected between the results obtained from the two windows selected to perform the blocks of simulations. This finding suggests that MDPT offer comparable performances regardless of the density of the points included in the set *T*. This is not surprising considering that the overall spatial extension of the points in the set *T* determines the range of variability of $\bar{d}$. Instead, depending on the tails of the tests, the correlation analysis indicated that the estimates of power were related either to the m number of points included in the subset I (for MDPTLT), or to the $\binom{n}{m}$ combinations of the subset I within the set T (for MDPT2T and MDPTRT). Since the power of a statistical test is generally positively correlated to the sample size, and provided that m and $\binom{n}{m}$ are quantities expressing the sample size, this finding is in agreement with theory, despite this theory has been developed for a few tests and mostly in a parametric framework (Acutis et al. 2012; Crawley 2013). Under a practical perspective, the MDPTLT seems to be endowed with the best performances in terms of power, also when m and $\binom{n}{m}$ are relatively small, while MDPT2T and MDPTRT appear to be more reliable when the ratio m/n tends towards the 50%. The estimates of type I error do not seem to be a criterion allowing to prefer one test to another according to the sampling size, as suggested by the almost complete lack of correlation with the above mentioned parameters. Despite the MC simulations were performed only for MDPT, they might be considered extendable to the corresponding MDRT, provided that B is large enough to achieve reliable estimates of p_e. In fact, as stated before, the randomization tests are unbiased approximations of their related permutation tests, whose accuracy can be improved up to the desired level (Ernst 2004).

The assessment of power and type I error through MC simulations is a numerical validation, since it is performed on known DGP. However, a biological validation is pivotal to verify the performances of a statistical test in the field (Thébaud et al. 2005). The biological validation was performed only on the MDRT in consideration of the above cited computational constraints. However, the 95% confidence intervals of p_r indicate a good level of accuracy and exclude ambiguity in the acceptance/rejection of the null hypotheses. Considering the combined results of the three MDRT, the points displaying a DR>0 within the network of sampling points covering the study area were clustered for *H. annosum* and randomly distributed for *H. irregulare*. Thus, for both fungal species, MDRT provided responses which were consistent among different tails and in agreement with the results obtained by Gonthier et al. (2012) by using spatial autocorrelation analyses, hence confirming the reliability of the MDRT in field conditions. Moreover, the advantage of performing the MDRT rather than autocorrelation analysis is intrinsic in the categorical measurement of the variable under investigation. The DR measured by Gonthier et al. (2012) required the counting of all fungal colonies of *Heterobasidion* spp. under a dissecting microscope, in addition to an appropriate sampling of colonies aimed at obtaining a large number of isolates (up to 40 per sampling point). The molecular analyses performed on these isolates were the last step to carry out the repartition of the DR between the two pathogenic species. This approach provided a quantitative information, which was essential to compare spores deposition between the two species as well as to carry out the autocorrelation analyses. However, the MDT could optimize the experimental design in similar trials. In fact, the assessment of the condition DR>0 could allow a less refined sampling procedure. For instance, molecular analyses could be dramatically reduced by pooling the samples of fungal mycelium of all isolates from each sampling point before DNA extraction. Also the number of isolates could be probably reduced without a substantial loss of information. Besides, the MDT could be performed on wide study areas, providing preliminary results to be further investigated turning to the quantitative level, but only in representative subareas.

The application of the MDT to the case study of the nut rot caused by *G. castanea* showed a possible way through which the designed geostatistical tests can be performed to gather information about a plant disease. Regardless of the area where the tests were performed, all MDT agreed in the identification of a random spatial pattern of the chestnut trees displaying the presence of *G. castanea* in at least one nut within the sampled trees. Since in half of the areas chestnuts were clustered, while in the other half they were not, it could be argued that the plantation density is not a variable influencing the spatial distribution of the pathogen. This conclusion seems to be confirmed by the absence of significant differences among the incidences of the pathogen among the areas. These findings suggest that the choice of the plantation density, which is a relevant issue for chestnut growers (Dong-Sheng et al. 2009), can be based on other parameters (e.g. yield productivity, intraspecific competition) rather than on the risk of transmission of *G. castanea* among neighbouring trees. This finding is relevant since, to date, very little was known about the relationship between the management practices and the incidence of *G. castanea*. However, it is important to stress that results from geostatistics do not replace biological and epidemiological investigations, but rather provide evidence about spatial distributions that can be helpful to formulate and to test hypotheses about disease dynamics. In the case of *G. castanea* further analyses are needed to determine the factors influencing the observed spatial patterns, since the infection pathways of *G. castanea* are still mainly unknown (Lione et al. 2014).

Despite the MDT approach is here proposed in the framework of plant pathology, if the assumption about the stochasticity of the processes under investigation are fulfilled, no constraints arise for its broader application in other research fields (e.g. ecology, forestry, economy). Even the number of spatial dimensions should not represent a substantial limit, since the one-dimensional case (e.g. plants in single-row alley) is a special case of the two-dimensional one (i.e. one coordinate is constant). The three-dimensional case could be included too, but it would require an extension of the MDT algorithms. Finally, the availability of accessible R algorithms and of a "point-and-click" software should facilitate the use of the MDT also among users lacking specific background in advanced statistics.

ACKNOWLEDGEMENTS

This study was partially supported by grants of Regione Piemonte through the activity of the Chestnut Growing Centre and of the University of Torino (60%).

LITERATURE CITED

Acutis, M., Scaglia, B., and Confalonieri, R. 2012. Perfunctory analysis of variance in agronomy, and its consequences in experimental results interpretation. Eur. J. Agron. 43:129-135.

Aldenderfer, M.S., and Blashfield, R.K. 1987. Cluster Analysis. SAGE Publications, London, UK.

Beeley, C. 2013. Web Application Development with R Using Shiny. Packt Publishing Ltd., Birmingham, UK.

Boots, B. 2003. Developing local measures of spatial association for categorical data. J. Geograph. Sys. 5:139-160.

Carsey, T.M., and Harden, J. J. 2014. Monte Carlo Simulation and Resampling Methods for Social Science. Sage Publications, Thousand Oaks, California, USA.

Chellemi, D.O., Rohrbach K.G., Yost R.S., and Sonoda R.M. 1988. Analysis of the spatial pattern of plant pathogens and diseased plants using geostatistics. Phytopathology 78:221-226.

Crawley, M.J. 2013. The R Book. Second Edition. John Wiley and Sons Ltd., Chichester, UK.

de Smith, M.J., Goodchild, M.F., and Longley, P. 2007. Geospatial Analysis: A Comprehensive Guide to Principles, Techniques And Software Tools. Troubador Publishing Ltd., Leicester, UK.

Dong-Sheng, Y., Ya-Li, H., Rui-Dong, T., Lin, Q., and Hong-Wen, H. 2009. The cultivation techniques of compactly planted chestnut (*Castanea mollissima* Bl.) for early fruiting and high yield. Acta Hort. 844:465.

Ernst, M. 2004. Permutation methods: a basis for exact inference. Statistical Science 19:676-685.

Ferrandino, F.J. 1998. Past nonrandomness and aggregation to spatial correlation: 2DCORR, a new approach for discrete data. Phytopathology 88:84-91.

Gonthier, P., Anselmi, N., Capretti, P., Bussotti, F., Feducci, M., Giordano, L., Honorati, T., Lione, G., Luchi N., Michelozzi, M., Paparatti, B., Sillo, F., Vettraino, A.M., and Garbelotto, M.

2014. An integrated approach to control the introduced forest pathogen *Heterobasidion irregulare* in Europe. Forestry 87:471-481.

Gonthier, P., Lione, G., Giordano, L., and Garbelotto, M. 2012. The American forest pathogen *Heterobasidion irregulare* colonizes unexpected habitats after its introduction in Italy. Ecol. Appl. 22:2135-2143.

Goslee, S.C. and Urban, D.L. 2007. The ecodist package for dissimilarity-based analysis of ecological data. J. Stat. Sofw. 22:1-19.

Gottwald, T.R., Avinent, L., Llácer, G., de Mendoza, A.H., and Cambra, M. 1995. Analysis of the spatial spread of sharka (plum pox virus) in apricot and peach orchards in eastern Spain. Plant Dis. 79:266-278.

Gray, S.M., Moyer, J.W., and Bloomfield, P. 1986. Two-dimensional distance class model for quantitative description of virus-infected plant distribution lattices. Phytopathology 76:243-248.

Hartwig, F.P. 2013. Two-tailed p-values calculation in permutation-based tests: a warning against “asymptotic bias” in randomized clinical trials. J. Clin. Trials 3: 145.

Jaime-Garcia, R., Trinidad-Correa, R., Felix-Gastelum, R., Orum, T.V., Wasmann, C.C., and Nelson, M.R. 2000. Temporal and spatial patterns of genetic structure of *Phytophthora infestans* from tomato and potato in the Del Fuerte Valley. Phytopathology 90:1188-1195.

Lamichhane, J.R., Fabi, A., Ridolfi, R., and Varvaro, L. 2013. Epidemiological study of hazelnut bacterial blight in central Italy by using laboratory analysis and geostatistics. PLoS ONE, 8(2), e56298. doi:10.1371/journal.pone.0056298.

Lione, G., Giordano, L., Sillo, F. and Gonthier, P. 2014. Testing and modelling the effects of climate on the incidence of the emergent nut rot agent of chestnut *Gnomoniopsis castanea*. Plant Pathol. doi: 10.1111/ppa.12319.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27:209-220.

Marchant, B.P., and Lark, R.M. 2004. Estimating variogram uncertainty. Math. Geol. 36:867-898.

Mitchell, A. 2009. The ESRI Guide To GIS Analysis. Volume 2. Spatial Measurements and Statistics. Environmental Systems Research Institute Press, Redlands, California, USA.

Nelson, S.C. 1995. STCLASS - spatiotemporal distance class analysis software for the personal computer. Plant Dis. 79:643-648.

Nelson, S.C., Marsh, P., and Campbell, C. 1992. 2DCLASS, a two-dimensional distance class analysis software for the personal computer. Plant Dis. 76:427-432.

Nelson, M.R., Orum T.V., Jaime-Garcia R., and Nadeem, A. 1999. Applications of geographic information systems and geostatistics in plant disease epidemiology and management. Plant Dis. 83:308-319.

Peres-Neto, P.R., and Olden, J.D. 2001. Assessing the robustness of randomization tests: examples from behavioural studies. Anim. Behav. 61:79-86.

Pizzato, J.A., Araújo, D.V., Galvanin, E.A., Júnior, J.R., Matos, Â.N., Vecchi, M., and Zavislak, F.D. 2014. Geostatistics as a methodology for studying the spatiotemporal dynamics of *Ramularia areola* in cotton crops. Am. J. Plant Sci. 5:2472-2479.

Thébaud, G., Peyrard, N., Dallot, S., Calonnec, A., and Labonne, G. 2005. Investigating disease spread between two assessment dates with permutation tests on a lattice. Phytopathology 95:1453-1461.

Visentin, I., Gentile, S., Valentino, D., Gonthier, P., Tamietti, G., and Cardinale F. 2012. *Gnomoniopsis castanea* sp. nov (Gnomoniaceae, Diaporthales) as the causal agent of nut rot in sweet chestnut. J. P. P. 94:411-419.

Webster, R., and Oliver, M.A. 2001. Geostatistics For Environmental Scientists. John Wiley and Sons Ltd., Chichester, UK.

Weidong, L. 2006. Transiogram: a spatial relationship measure for categorical data. Int. J. Geograph. Inf. Sci. 20:693-699.

Table 1

Test type	**Test**	**Tail**	**Null hypothesis H_0**	**Input**	**Output**
Permutation	MDPT2T	2-tailed	the spatial pattern of level γ is random	- γ: level assigned to points in I - x and y: coordinates of points in T - α: significance level cut-off	- $\bar{d}_0$: observed mean value of the triangular Euclidean distance matrix among the points in I - $\overline{D}$: mean of the permutation distribution - p_e: exact p-value
	MDPTLT	left-tailed	the spatial pattern of level γ is not clustered		
	MDPTRT	right-tailed	the spatial pattern of level γ is not dispersed		
Randomization	MDRT2T	2-tailed	the spatial pattern of level γ is random	- γ: level assigned to points in I - x and y: coordinates of points in T - α: significance level cut-off - B: number of random combinations - λ: confidence level for the p-value	$\bar{d}_0$: observed mean value of the triangular Euclidean distance matrix among the points in I - $\overline{D}$: mean of the randomization distribution - p_r: randomization p-value - L_{pr}: lower bound of the λ confidence interval of p_r - U_{pr}: upper bound of the λ confidence interval of p_r
	MDRTLT	left-tailed	the spatial pattern of level γ is not clustered		
	MDRTRT	right-tailed	the spatial pattern of level γ is not dispersed		

Table 2

Test	**DGP verifying H_0**	**DGP not verifying H_0**	**Number of simulations per DGP to estimate power within each block**	**Number of simulations per DGP to estimate type I error within each block**
MDPT2T	PP1	PP2; PP3	$5 \cdot 10^3$ PP2 + $5 \cdot 10^3$ PP3	$1 \cdot 10^4$ PP1
MDPTLT	PP1; PP3	PP2	$1 \cdot 10^4$ PP2	$5 \cdot 10^3$ PP1 + $5 \cdot 10^3$ PP3
MDPTRT	PP1; PP2	PP3	$1 \cdot 10^4$ PP3	$5 \cdot 10^3$ PP1 + $5 \cdot 10^3$ PP2

Table 3.

		MDPT2T				**MDPTLT**				**MDPTRT**			
		6 × 6 units window		**4 × 4 units window**		**6 × 6 units window**		**4 × 4 units window**		**6 × 6 units window**		**4 × 4 units window**	
m	$\binom{n}{m}$	**power**	**type I error**	**power**	**type I error**	**power**	**type I error**	**power**	**type I error**	**power**	**type I error**	**power**	**type I error**
2	105	0.7638	0.0499	0.7675	0.0419	0.8983	0.0264	0.9006	0.0246	0.6528	0.0237	0.6503	0.0239
3	455	0.8768	0.0491	0.8744	0.0496	0.9998	0.0241	0.9997	0.0256	0.7476	0.0229	0.7470	0.0240
4	1,365	0.9312	0.0499	0.9275	0.0471	1.0000	0.0232	1.0000	0.0250	0.8844	0.0251	0.8814	0.0239
5	3,003	0.8897	0.0491	0.8911	0.0507	1.0000	0.0256	1.0000	0.0254	0.9054	0.0260	0.9071	0.0248
6	5,005	0.9528	0.0524	0.9487	0.0543	1.0000	0.0241	1.0000	0.0272	0.9355	0.0255	0.9320	0.0274
7	6,435	0.9513	0.0484	0.9526	0.0458	1.0000	0.0245	1.0000	0.0256	0.9562	0.0248	0.9566	0.0248
8	6,435	0.9569	0.0517	0.9537	0.0504	1.0000	0.0259	1.0000	0.0244	0.9619	0.0266	0.9654	0.0254
9	5,005	0.9561	0.0473	0.9567	0.0516	1.0000	0.0247	1.0000	0.0250	0.9594	0.0247	0.9633	0.0235
10	3,003	0.9482	0.0471	0.9517	0.0482	1.0000	0.0260	1.0000	0.0247	0.9532	0.0261	0.9491	0.0248
11	1,365	0.9387	0.0484	0.9367	0.0487	1.0000	0.0242	1.0000	0.0225	0.9345	0.0255	0.9355	0.0257
12	455	0.9040	0.0488	0.9036	0.0511	1.0000	0.0259	1.0000	0.0250	0.9171	0.0245	0.9160	0.0254
13	105	0.8262	0.0531	0.8267	0.0495	1.0000	0.0257	1.0000	0.0231	0.8530	0.0245	0.8588	0.0235
average		0.9080	0.0496	0.9075	0.0491	0.9915	0.0250	0.9917	0.0248	0.8884	0.0250	0.8885	0.0247
$\rho(m)$		0.2168	-0.2039	0.2587	0.2587	0.6504*	0.2767	0.6504*	-0.5149	0.3846	0.1754	0.4196	0.1343
$\rho(m)$ **p-value**		0.4991	0.5251	0.4169	0.4169	0.0220	0.3839	0.0220	0.0867	0.2184	0.5855	0.1766	0.6774
$\rho\binom{n}{m}$		0.9046*	-0.2487	0.8905*	0.2686	0.5324	-0.2053	0.5324	0.3611	0.8905*	0.6738*	0.8622*	0.4143
$\rho\binom{n}{m}$ **p-value**		0.0001	0.4358	0.0001	0.3987	0.0747	0.5221	0.0747	0.2489	0.0001	0.0163	0.0003	0.1806

Table 4.

		Test								
		MDPT2T	**MDPTLT**	**MDPTRT**	**MDRT2T** **B=10^2**	**MDRTLT** **B=10^2**	**MDRTRT** **B=10^2**	**MDRT2T** **B=5·10^2**	**MDRTLT** **B=5·10^2**	**MDRTRT** **B=5·10^2**
Area	**C1** $\bar{d}_0$=21.2 m $\begin{pmatrix} n_{C1} \\ m_{C1} \end{pmatrix} = \begin{pmatrix} 14 \\ 10 \end{pmatrix} = 1{,}001$	$\bar{D}$=19.9 m p_e=0.301	$\bar{D}$=19.9 m p_e=0.856	$\bar{D}$=19.9 m p_e=0.145	$\bar{D}$=19.9 m p_r=0.29 L_{pr}=0.26 U_{pr}=0.40	$\bar{D}$=19.6 m p_r=0.84 L_{pr}=0.78 U_{pr}=0.89	$\bar{D}$=19.7 m p_r=0.15 L_{pr}=0.09 U_{pr}=0.25	$\bar{D}$ = 19.8 m p_r=0.31 L_{pr}=0.26 U_{pr}=0.32	$\bar{D}$ = 20.0 m p_r=0.85 L_{pr}=0.83 U_{pr}=0.88	$\bar{D}$ = 20.0 m p_r=0.14 L_{pr}=0.13 U_{pr}=0.18
	C2 $\bar{d}_0$=22.3 m $\begin{pmatrix} n_{C2} \\ m_{C2} \end{pmatrix} = \begin{pmatrix} 17 \\ 9 \end{pmatrix} = 24{,}310$	$\bar{D}$=21.2 m p_e=0.5355	$\bar{D}$=21.2 m p_e=0.7240	$\bar{D}$=21.2 m p_e=0.2760	$\bar{D}$=21.1 m p_r=0.47 L_{pr}=0.46 U_{pr}=0.62	$\bar{D}$=21.3 m p_r=0.68 L_{pr}=0.63 U_{pr}=0.83	$\bar{D}$=21.1 m p_r=0.21 L_{pr}=0.19 U_{pr}=0.46	$\bar{D}$=21.2 m p_r=0.59 L_{pr}=0.48 U_{pr}=0.60	$\bar{D}$ = 21.2 m p_r=0.73 L_{pr}=0.68 U_{pr}=0.78	$\bar{D}$ = m p_r=0.26 L_{pr}=0.24 U_{pr}=0.35
	NC1 $\bar{d}_0$=18.8 m $\begin{pmatrix} n_{NC1} \\ m_{NC1} \end{pmatrix} = \begin{pmatrix} 14 \\ 8 \end{pmatrix} = 3{,}003$	$\bar{D}$=21.5 m p_e=0.158	$\bar{D}$ = 21.5 m p_e=0.088	$\bar{D}$ = 21.5 m p_e=0.913	$\bar{D}$=21.4 m p_r=0.16 L_{pr}=0.09 U_{pr}=0.21	$\bar{D}$=21.7 m p_r=0.13 L_{pr}=0.04 U_{pr}=0.21	$\bar{D}$=21.8 m p_r=0.90 L_{pr}=0.86 U_{pr}=0.96	$\bar{D}$ = 21.5 m p_r=0.16 L_{pr}=0.14 U_{pr}=0.19	$\bar{D}$ = 21.5 m p_r=0.08 L_{pr}=0.07 U_{pr}=0.11	$\bar{D}$ =m p_r=0.90 L_{pr}=0.88 U_{pr}=0.92
	NC2 $\bar{d}_0$=32.7 m	$\bar{D}$=31.6 m p_e=0.6534	$\bar{D}$=31.6 m p_e=0.6509	$\bar{D}$=31.6 m p_e=0.3491	$\bar{D}$=31.7 m p_r=0.63 L_{pr}=0.43	$\bar{D}$=31.5 m p_r=0.58 L_{pr}=0.53	$\bar{D}$=31.5 m p_r=0.33 L_{pr}=0.31	$\bar{D}$ = 31.6 m p_r=0.60 L_{pr}=0.59	$\bar{D}$ = 31.8 m p_r=0.65 L_{pr}=0.60	$\bar{D}$ = 31.6 m p_r=0.35 L_{pr}=0.29 U_{pr}=0.41

$$\binom{n_{NC2}}{m_{NC2}} = \binom{17}{11} = 12{,}376$$

U_{pr}=0.84 U_{pr}=0.72 U_{pr}=0.40 U_{pr}=0.65 U_{pr}=0.71

585

Table 1. For each test included in the Mean Distance Tests (MDT) the tail, the null hypothesis, the input required and the output provided are indicated. Tests are divided according to the underlying resampling technique (test type) and identified by an acronym (test).

Table 2. Data generating processes (DGPs) verifying or not verifying the null hypothesis H_0 of each test included in Mean Distance Permutation Tests (MDPT) and combinations of the three DGPs used to perform the blocks of Monte Carlo simulations for power and type I error estimation.

Table 3. Estimates of power and type I error for the Mean Distance Permutation Tests (MDPT) obtained through Monte Carlo simulations and results of the correlation analysis. The estimates are provided for each block of simulations ranked according to the *m* values and divided for two-tailed, left-tailed and right-tailed tests (MDPT2T, MDPTRT, and MDPTLT) and window size. The number of combinations $\binom{n}{m}$ enumerated for each value of *m* is listed. The average of power and type I error as well as the Spearman correlation coefficient between the estimates and *m* [i.e. $\rho(m)$] and $\binom{n}{m}$ [i.e. $\rho\binom{n}{m}$] are reported with the related p-value for all tests and window sizes. The symbol * indicates correlation coefficients significant at 0.05 cut-off.

Table 4. Output of the Mean Distance Tests for areas C1, C2, NC1 and NC2. The output includes the mean value $\overline{D}$ of the probability mass function (PMF), the exact p-value (p_e) for permutation tests, the randomization p-value (p_r) with lower (L_{pr}) and upper (U_{pr}) bounds of its 95% confidence interval. For randomization tests the output is divided according to the number *B* of combinations randomly selected to perform the tests. The observed mean value of the triangular Euclidean

distance matrix among the m out of n chestnut trees carrying at least one infected nut ($\bar{d}_0$) and the number of possible combinations $\binom{n}{m}$ are reported for each area.

Fig. 1. Maps of the sampling points in the Circeo National Park that displayed the presence of spores of *Heterobasidion annosum* (A) and *Heterobasidion irregulare* (B), defining the subsets I_1 and I_2 respectively.

Fig. 2. Maps of chestnut trees of the “Vivaio Gambarello” orchard carrying at least one nut infected by *Gnomoniopsis castanea* (level γ) in areas C1 (A), C2 (B), NC1 (C) and NC2 (D).

FIG. 1

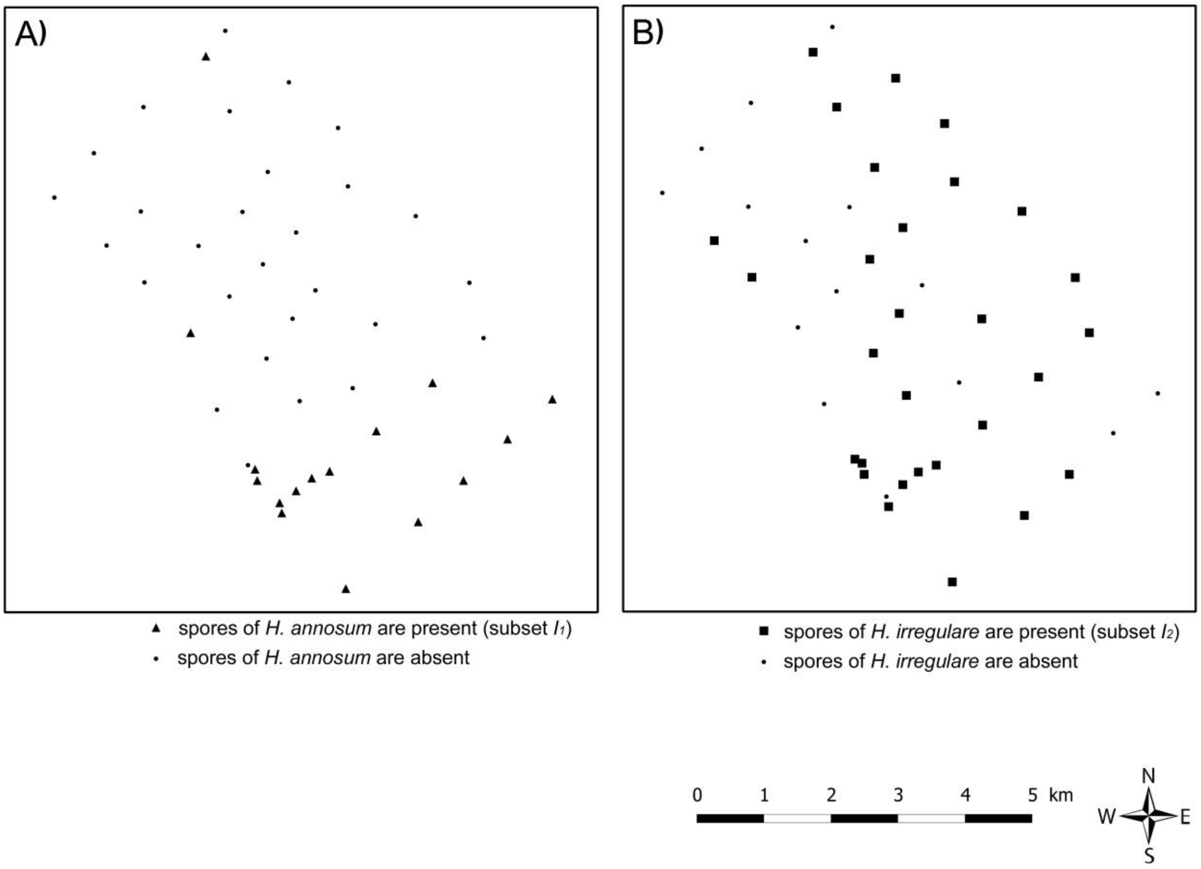
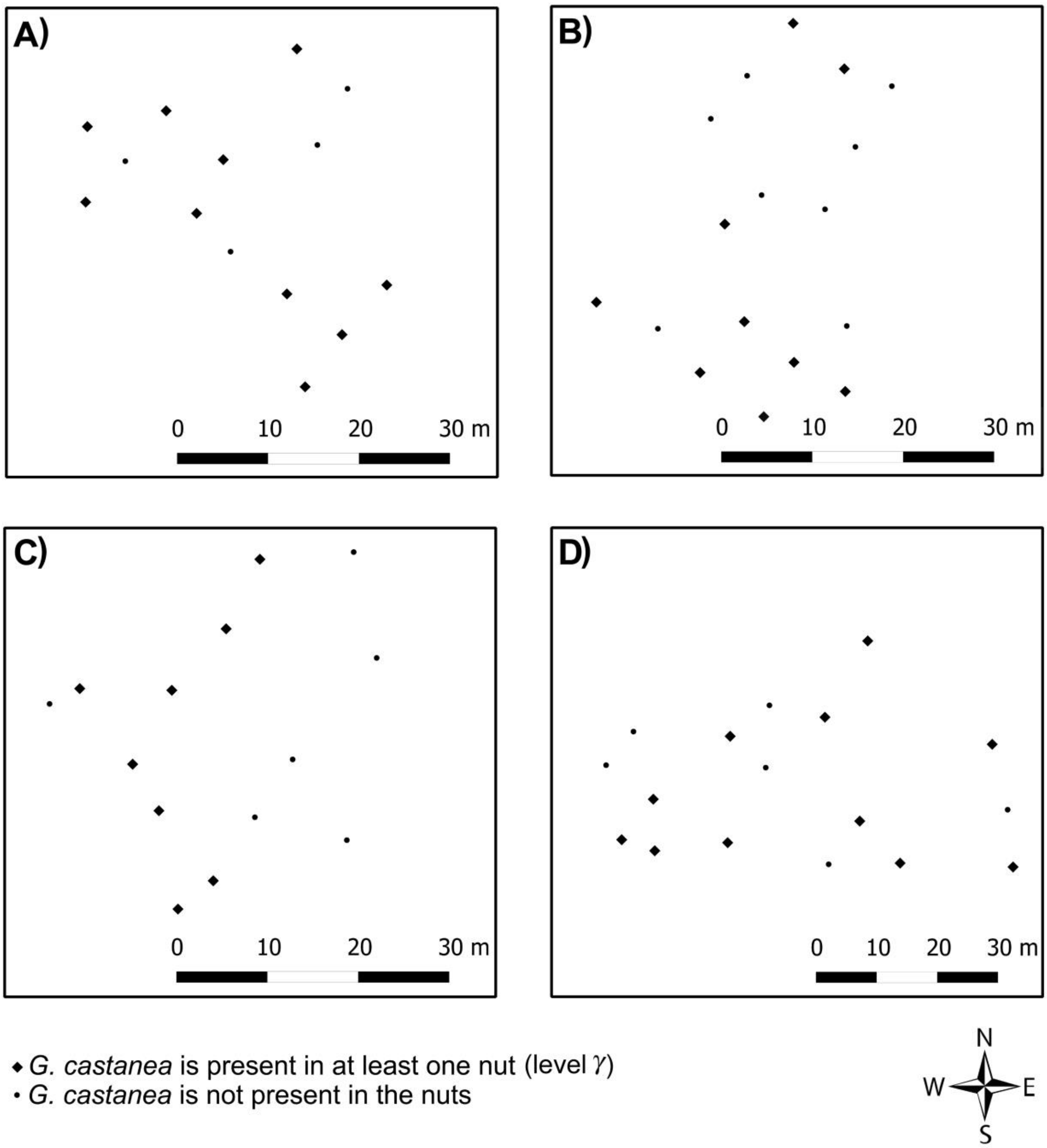

A)
B)
▲ spores of H. annosum are present (subset I1)
• spores of H. annosum are absent
■ spores of H. irregulare are present (subset I2)
• spores of H. irregulare are absent
0
1
2
3
4
5
km
N
W
E
S

FIG. 2


A)
0
10
20
30 m
B)
0
10
20
30 m
C)
0
10
20
30 m
D)
0
10
20
30 m
◆ G. castanea is present in at least one nut (level γ)
• G. castanea is not present in the nuts
N
W
E
S