

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

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

This is a pre print version of the following article:

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/1552090> since 2016-02-01T17:02:18Z

Published version:

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

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)



UNIVERSITÀ DEGLI STUDI DI TORINO

1
2
3
4
5
6
7
8
9
10
11
12
13

This is an author version of the contribution:

Questa è la versione dell'autore dell'opera:

*[Lione G., Gonthier P., 2016. *Phytopathology*, 106, pp. 19–28,*

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

The definitive version is available at:

La versione definitiva è disponibile alla URL:

[<http://apsjournals.apsnet.org/doi/10.1094/PHYTO-05-15-0112-R>]

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

15

16 G. Lione, and P. Gonthier

17

18 Department of Agricultural, Forest and Food Sciences, University of Torino, Largo P. Braccini 2,
19 10095 Grugliasco, Italy.

20

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

22

23 **ABSTRACT**

24

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

27

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

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

42

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

45

46 INTRODUCTION

47

48 Analyzing the spatial pattern of plant diseases may be pivotal to elucidate the ecology, the
49 epidemiology and the infection biology of pathogens as well as the mechanisms underlying host-
50 pathogen interactions and the spread of epidemics (Nelson et al. 1999). A large body of literature
51 deals with the application of Geographic Information Systems (GIS) in conjunction with statistical
52 and geostatistical methods to investigate peculiar traits of plants diseases, to test biologically
53 relevant hypotheses and to build predictive and/or explicative models (Nelson et al. 1999).

54 Examples of GIS and geostatistical applications can be found in both agriculture and forestry on a
55 broad range of diseases, hosts and pathogens, including viruses, bacteria and fungi. For instance,
56 GIS and geostatistical analyses were used to relate the presence of tomato virus vectors to the
57 spatial pattern of the symptoms in tomato (*Solanum lycopersicum* L.) crops (Nelson et al. 1999).

58 Analogous analyses were performed to test the association between genetic variations in cotton leaf
59 curl viruses and the disease severity in *Gossypium* spp. fields (Nelson et al. 1999) and to investigate
60 the dispersion mechanisms of the plum pox potyvirus in orchards of *Prunus armeniaca* L. and *P.*

61 *persica* (L.) Batsch (Gottwald et al. 1995). Similar approaches were carried out to elucidate the role
62 of pedoclimatic factors on the incidence of the bacterial blight caused by *Xanthomonas arboricola*
63 *pv. corylina* on *Corylus avellana* L. (Lamichhane et al. 2013). GIS and geostatistics were also used

64 to explore the spatial distribution of genotypes of *Phytophthora infestans* (Mont.) de Bary in
65 orchards of *S. lycopersicum* and *Solanum tuberosum* L. affected by late blight disease (Jaime-

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

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

83 Several methods are available to assess whether the spatial distribution of points is clustered,
84 random or dispersed, including the Nearest Neighbor Index (NNI), the Ripley's K function and the
85 Nearest Neighbor Hierarchical Clustering (NNHC), whose significance is generally estimated with
86 Monte Carlo (MC) simulations (Mitchell 2009). The rationale of MC simulations lies in the
87 comparison between the observed points location and the location of a large number of points
88 samples drawn from a predefined data generating process (DGP) known as point process (Crawley
89 2013; Carsey and Harden 2014). The choice of the appropriate point process depends upon the null
90 hypothesis being tested (de Smith et al. 2007).

91 The spatial distribution of the quantitative variable associated with points is generally
92 assessed through spatial autocorrelation analyses involving the Mantel test, the estimation of
93 variograms and the calculation of autocorrelation indexes such as the Geary's c , the Moran's I and
94 the Getis-Ord general G -statistic at global or local scale (Mantel 1967; Mitchell 2009; Webster and
95 Oliver 2001). To account for the stochastic uncertainty related to these methods, asymptotic theory
96 and heuristic procedures are available (Goslee and Urban 2007; Marchant and Lark 2004; Mitchell
97 2009). While the above cited techniques are routinely applied and embedded in some major GIS
98 and statistical software (Mitchell 2009), the spatial distribution of a categorical variable associated
99 with points is still a topic of active research and ongoing development. In the last decades plant
100 pathologists have proposed and validated some conceptual and technical solutions to this issue. For
101 instance, the software package 2DCLASS was designed to perform the Gray's analysis aimed at
102 detecting the spatial pattern of plant diseases (Gray et al. 1986; Nelson et al. 1992). 2DCLASS was
103 further improved by the STCLASS package (Nelson 1995) and by a MC-based approach to
104 investigate the spatiotemporal pattern of the spread of epidemics (Thébaud et al. 2005). A
105 correlation-based technique was also proposed to detect the spatial distribution of discrete data
106 through the 2DCORR package (Ferrandino 1997). More recently, an extension of local measures of
107 spatial association was suggested to deal with the same kind of data (Boots 2003). The above cited
108 solutions were designed to analyze binomial categorical data (e.g. infected/healthy plant) in lattices,
109 where points were approximated to cells in a regular grid, including missing points (e.g. missing
110 plants). While this approximation is suitable to model many field conditions where plants are
111 located in the space according to a predefined geometric pattern, like in nurseries, in orchards and in
112 regular plantations, no application to forestry, to irregular plantations and to natural seedlings
113 regeneration has been reported so far. Despite transiogram analyses were proposed to overcome the
114 constraints related to the plants plantation scheme, the discrepancy between experimental
115 transiograms and idealized ones can occur, affecting the interpretation of the results (Weidong
116 2006).

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

121

122 MATERIALS AND METHODS

123

124 **Overview and software design.** Let $T = \{t_1, t_2, \dots, t_n\}$ be a finite set of n points with known
 125 x and y coordinates in a Cartesian plane and let $I \subset T$ be a subset of T including m ($2 \leq m \leq n - 2$)
 126 points. For instance, the points in the set T could be plants and the points in the subset I could be the
 127 plants infected by some pathogens. In other terms, the m points in the subset I are those points of
 128 the set T which a level γ (i.e. “infected”) of a categorical variable Γ (e.g. “health status”) has been
 129 assigned to. Let \bar{d} be an overall index of the distances that separate m points in a plane, calculated

130 as the mean of the values stored in the $m \times m$ triangular Euclidean distance matrix of the points. Let
 131 \bar{d}_0 be the observed value of \bar{d} , which is calculated for the m points included in the subset I . Finally,

132 let be $\binom{n}{m}$ a binomial coefficient, representing the number of possible arrangements of m elements
 133 drawn from a set of n elements. Within the permutation tests framework, the probability mass

134 function (PMF) of \bar{d} is obtained by calculating \bar{d} for each i^{th} combination $\left[1 \leq i \leq \binom{n}{m} \right]$ through

135 which m points of the set T can be randomly assigned to the subset I (Carsey and Harden 2014).

136 Instead, within the randomization tests framework, the PMF is estimated by calculating \bar{d} on a

137 random sample without replacement of B combinations $\left[1 < B < \binom{n}{m} \right]$ (Carsey and Harden 2014).

138 The main core of this work is to determine from the PMF, with a predefined significance level cut-

139 off α , whether \bar{d}_0 is either significantly lower (i.e. located towards the left tail) or higher (i.e.
 140 located towards the right tail) than expected under the random assignment of γ (i.e. random
 141 definition of the subset I within the set T). The first case indicates a clustered spatial pattern of the
 142 level γ , while the second occurs in a dispersed spatial pattern of the same level. This is equivalent to
 143 test if the infected plants are nearer or further apart than expected according to a random
 144 distribution of the infected plants within the sampled plants. To deal with this issue the Mean
 145 Distance Tests (MDT) approach is proposed here.

146 MDT are based on the assumption that the x and y coordinates of points in the set T are fixed
 147 and that only the assignment of the level γ is a stochastic process. The MDT consist of 3
 148 permutation tests (Mean Distance Permutation Tests - MDPT) and 3 randomization tests (Mean
 149 Distance Randomization Tests - MDRT). Both permutation and randomization tests are divided
 150 according to the tails of the PMF they refer to (Hartwig 2013). MDPT2T is the two-tailed (2T)
 151 permutation test, MDPTLT the left-tailed (LT) and MDPTRT the right-tailed (RT), respectively.
 152 Similarly, the MDRT are designed in the two-tailed version (MDRT2T), in the left-tailed
 153 (MDRTLTL) and in the right-tailed (MDRTRT) ones (Table 1). Once the above described steps to
 154 obtain the PMF and to calculate \bar{d}_0 are performed, the mean value \bar{D} of the PMF is calculated, the
 155 exact p-value (p_e) is determined for MDPT and the randomization p-value (p_r) is determined for
 156 MDRT as reported in Carsey and Harden (2014) and Ernst (2004). The adequacy of the number B
 157 selected to perform the MDRT is assessed by calculating the lower (L_{pr}) and upper (U_{pr}) bounds of
 158 the confidence interval for p_r at user-defined level λ (e.g. 0.95). The confidence interval is
 159 calculated from the binomial distribution as described in Ernst (2004). Whenever the condition
 160 $L_{pr} \leq \alpha \leq U_{pr}$ is verified, p_r is deemed to be ambiguous and B is increased until the sampling
 161 adequacy is achieved and, thus, ambiguity is solved (Ernst 2004).

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

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

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

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

197 **Biological validation.** The MDRT were validated on data gathered from Gonthier et al.
 198 (2012). In this study, 44 sampling points equipped with spores trapping devices were located within
 199 a 3030 ha forest in the Circeo National Park, in central Italy. Spore trapping devices allowed to
 200 determine the spores deposition rate (DR), expressed as the number of viable spores per squared
 201 meter per hour (spores \cdot m⁻² \cdot h⁻¹), of two fungal pathogens causing root rot on conifers. The first
 202 pathogen, *Heterobasidion annosum* (Fr.) Bref., is native in the area, while the second one, *H.*
 203 *irregulare*, is an alien invasive species. Geostatistical analyses of spatial autocorrelation performed
 204 on the DR showed that *H. irregulare* was ubiquitous and distributed in the area according to a
 205 random spatial pattern, while *H. annosum* showed significant clustering around patches of conifers.

206 To validate the MDRT, the set T was defined including all $n=44$ sampling points. Two
 207 categorical variables Γ_1 (i.e. “presence of *H. annosum* spores”) and Γ_2 (i.e. “presence of *H.*
 208 *irregulare* spores”) were defined. For Γ_1 the γ_1 level (i.e. “*H. annosum* spores are present”) was
 209 assigned to the m_1 sampling points with *H. annosum* DR>0, which were included in the subset I_1 .
 210 Similarly, the γ_2 level (i.e. “*H. irregulare* spores are present”) was assigned to the m_2 sampling

211 points with *H. irregulare* DR>0 to define the subset I_2 . MDRT2T, MDRTLTL and MDRTTRT with
 212 $\alpha=0.05$, $B=10^4$ and $\lambda=0.95$ were performed on both γ_1 and γ_2 levels.

213 **Application to a case study.** An application of the MDT to a case study was carried out to
 214 test the relation between the plantation density and the incidence of *Gnomoniopsis castanea*, an
 215 emerging fungal pathogen causing the nut rot of chestnut (Visentin et al. 2012). During October
 216 2013, the coordinates of 203 sweet chestnuts (*C. sativa*) were recorded in UTM WGS84 zone 32 N
 217 system (m) with a GPS device (Magellan Mobile Mapper 6, Magellan Navigation Inc., Santa Clara,
 218 CA, USA). The trees grew in the sweet chestnut orchard “Vivaio Gambarello”, set in the north-west
 219 of Italy (E 394,925; N 4,906,885). A NNHC analysis (Mitchell 2009) was performed on CrimeStat
 220 3.3. (Ned Levine & Associates, Houston, TX, USA) with $2 \cdot 10^3$ iterations and significance level cut-
 221 off set to 0.05. The two clusters of sweet chestnuts including the largest number of trees (areas C1
 222 and C2, see results) were selected and two not clustering groups (areas NC1 and NC2) with the
 223 same number of sweet chestnuts were randomly chosen. The mean value of the triangular Euclidean
 224 distance matrix among all the sweet chestnuts was calculated for areas C1, C2, NC1 and NC2. Up
 225 to 40 nuts per tree were collected from the crown of each sweet chestnut in the above mentioned
 226 areas. Fragments of the nuts kernel were plated in Petri dishes on Malt Extract Agar (MEA) to
 227 assess the presence/absence of *G. castanea* in the fruit tissues at the tree level. Isolations and fungal
 228 identification were performed as described by Lione et al. (2014). The incidence of *G. castanea* was
 229 calculated as the ratio, in percent, between the m_{C1}, m_{C2}, m_{NC1} and m_{NC2} trees carrying at least one
 230 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}
 231 and n_{NC2} trees growing in each area (i.e. sets T_{C1}, T_{C2}, T_{NC1} and T_{NC2}). The categorical variable F
 232 (i.e. “presence of *G. castanea* in at least one nut”) was defined and the level γ (i.e. “*G. castanea* is
 233 present in at least one nut”) was assigned to the m_{C1}, m_{C2}, m_{NC1} and m_{NC2} trees. The incidence of the
 234 pathogen was compared among the four above mentioned areas with a χ^2 test performed with a
 235 significance cut-off of 0.05. For each area \bar{d}_0 and $\binom{n}{m}$ were calculated. MDRT2T, MDRTLTL and

236 MDRTRT with $\alpha=0.05$ and MDRT2T, MDRTLTLT and MDRTRT with $\alpha=0.05$, $B=10^2$, $B=5 \cdot 10^2$ and
 237 $\lambda=0.95$ were performed on the γ level for every area.

238

239 **RESULTS**

240

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

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

259 ($P > 0.05$) were observed between the estimates of type I error and either m or $\binom{n}{m}$, with the

260 exception of MDPTRT in the 6×6 units window (Table 3).

261 **Biological validation.** For the variable Γ_1 , the level γ_1 was assigned to $m_1=16$ sampling
 262 points that fulfilled the condition *H. annosum* DR>0, defining the subset I_1 (Fig. 1A). For γ_1 , the
 263 value of \bar{d}_0 attained 2767 m, while \bar{D} was 3449 m in MDRT2T and 3443 m in both MDRTLTLT and
 264 MDRTTRT. Based on MDRT2T, sampling points where spores of *H. annosum* had been detected
 265 were not randomly distributed within the sampling points ($p_r=0.0122$, $L_{pr}=0.0117$, $U_{pr}=0.0158$).
 266 MDRTLTLT indicated a clustered spatial pattern of the points with *H. annosum* DR>0 within the
 267 sampling points ($p_r=0.0092$, $L_{pr}=0.0065$, $U_{pr}=0.0113$). Finally, MDRTTRT was not significant,
 268 showing a not dispersed spatial distribution of the points with *H. annosum* DR>0 within the
 269 sampling points ($p_r=0.9892$, $L_{pr}=0.9875$, $U_{pr}=0.9918$). The subset I_2 was defined by assigning the
 270 level γ_2 of the variable Γ_2 to the $m_2=29$ points that satisfied the condition *H. irregulare* DR>0 (Fig.
 271 1B). In this case, \bar{d}_0 attained a value of 3281 m, while \bar{D} ranged from 3445 m in MDRTLTLT to 3446
 272 m in MDRT2T and MDRTTRT. MDRT2T output indicated that sampling points where spores of *H.*
 273 *irregulare* had been identified were randomly distributed within the sampling points ($p_r=0.2554$,
 274 $L_{pr}=0.2422$, $U_{pr}=0.2699$). According to MDRTLTLT, points with *H. irregulare* DR>0 were not
 275 clustered within the sampling points ($p_r=0.1278$, $L_{pr}=0.1272$, $U_{pr}=0.1402$), while the MDRTTRT
 276 showed a not dispersed spatial pattern for the same points ($p_r=0.8739$, $L_{pr}=0.8636$, $U_{pr}=0.8781$). In
 277 all MDRT performed the condition $L_{pr} \leq \alpha \leq U_{pr}$ was not verified for $B=10^4$.

278 **Application to the case study.** The NNHC showed the presence of 24 first order clusters,
 279 comprising two to five trees, and two second order clusters (areas C1 and C2), composed by four
 280 and five first order clusters with a total of $n_{C1}=14$ and $n_{C2}=17$ sweet chestnuts, respectively ($P < 0.05$)
 281 (Fig. 2A and 2B). The same number of trees was used to define the areas NC1 ($n_{NC1}=14$) and NC2
 282 ($n_{NC2}=17$) (Fig. 2C and 2D). The mean value of the triangular Euclidean distance matrix among all
 283 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

284 assigned to the $m_{C1}=10, m_{C2}=9, m_{NC1}=8$ and $m_{NC2}=11$ sweet chestnuts carrying at least one nut
 285 infected by *G. castanea* (Fig. 2). The incidence of *G. castanea* was 71.4% in C1, 52.9% in C2,
 286 57.1% in NC1 and 64.7% in NC2. The χ^2 test indicated no significant differences among the
 287 incidence level of the four areas ($P=0.7312$). The \bar{d}_0 distance ranged from 18.8 m to 32.7 m, with
 288 the lowest values observed in C1 and C2, while $\binom{n}{m}$ was comprised between 1,001 and 24,310,
 289 depending on the area. The MDT performed were never significant ($p_e > 0.05; p_r > 0.05$), regardless of
 290 the area, indicating a random (2T), not clustered (LT) and not dispersed (RT) spatial distribution of
 291 sweet chestnuts infected by *G. castanea* within the sampled trees. The B values were adequate to
 292 perform the MDRT since the condition $L_{pr} \leq \alpha \leq U_{pr}$ was not verified, with the exception of the
 293 MDRTLTL carried out in NC1 for $B=10^2$. Increasing B values reduced the width of the interval [$L_{pr},$
 294 U_{pr}] for every MDRT in all areas (Table 4).

295

296 **DISCUSSION**

297

298 The analysis of the spatial pattern of plant diseases is a pivotal issue in plant pathology since
 299 it is aimed at gathering relevant information about biological, epidemiological and ecological
 300 aspects of pathogens. In this regard, during the last decades, an increasing interest has been
 301 addressed by plant pathologists to the development and the use of statistical and geostatistical
 302 methods. It is worth noting that the majority of these methods was mainly designed to analyze
 303 specific kinds of variables in a limited range of field conditions. A large body of literature dealt
 304 with the spatial distribution of relevant phytopathological measures on the continuous or ordinal
 305 scale, while few studies were focused on the spatial pattern of categorical variables. Moreover,
 306 many researches carried out on categorical variables proposed geostatistical methods aimed at
 307 analyzing diseases in lattices and in regular plantations. The application of such methods often
 308 requires the user to own a solid background in mathematics, advanced statistics and information

309 technology, since the algorithms performing the tests are rarely wrapped into a user-friendly “point-
310 and-click” interface. These aspects may thwart the diffusion of some statistical and geostatistical
311 tests in phytopathology, despite they were designed explicitly to analyze plant diseases. Within this
312 framework, the main goal of our study was to propose the MDT as a series of geostatistical tests to
313 assess the spatial pattern of plant diseases when the variable of phytopathological interest is
314 categorical and to provide the user with an intuitive “point-and-click” software to perform the tests.

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

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

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

342 The algorithms proposed for the MDT are largely based on the estimation of the PMF of the
343 distance parameter \bar{d} through either permutation or randomization. Both permutation and
344 randomization are currently considered robust and flexible standards for the assessment of the PMF
345 of parameters lacking a solid distributional theory (Carsey and Harden 2014; Ernst 2004; Peres-
346 Neto and Olden 2001). Whenever possible, the permutation approach should be preferred, since the
347 randomization leads to an estimate of the permutation results, implying a higher degree of
348 uncertainty in the response. However, permutation methods may pose heavy computational issues
349 in terms of time consumption and technical feasibility (Ernst 2004). Combinatorics shows that, even
350 for moderate sample sizes, the amount of data generated during a permutation test may be
351 extremely large, requiring an excessively long time to be processed, or even exceeding the available
352 computational power of the computer. Thus, the limits of the computer performances may impose
353 the switch from the permutation to the randomization approach (Carsey and Harden 2014). This
354 switch implies a cost in terms of uncertainty, that in the case of the MDRT affects the value of p_r .
355 To deal with this issue, the calculation of confidence intervals for p_r were embedded in the MDRT
356 algorithms as indicated by Ernst (2004). It is worth noting that the theory of resampling methods
357 suggests that a higher accuracy in the results of randomization may be acquired by increasing the
358 number of combinations randomly selected to perform the test (Carsey and Harden 2014; Ernst
359 2004). This is remarkably relevant when the randomization p-value tends to approach α , the cut-off
360 level dividing the regions of acceptance/rejection of the null hypothesis under the estimated PMF.

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

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

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

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

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

399 The assessment of power and type I error of permutation tests requires an heuristic
400 approach based on MC simulations (Peres-Neto and Olden 2001; Thébaud et al. 2005). The average
401 and the single values obtained for power and type I error estimates of MDPT were in agreement
402 with those reported for analogous geostatistical tests by Thébaud et al. (2005). On average the
403 power of both two-tailed and one-tailed tests was larger than 0.80, while the type I error was lower
404 than 0.05, as generally recommended to ensure the trustworthiness of statistical tests (Crawley
405 2013). The number of simulations performed within each block and the number of blocks were
406 deemed to be largely sufficient to provide reliable estimates of the power and the type I error, in
407 agreement with previously reported data (Carsey and Harden 2014; Ernst 2004; Thébaud et al.
408 2005). The window sizes seemed not to be influential on the estimates of the power and of the type
409 I error, as demonstrated by the small differences detected between the results obtained from the two
410 windows selected to perform the blocks of simulations. This finding suggests that MDPT offer
411 comparable performances regardless of the density of the points included in the set T . This is not
412 surprising considering that the overall spatial extension of the points in the set T determines the

413 range of variability of \bar{d} . Instead, depending on the tails of the tests, the correlation analysis
414 indicated that the estimates of power were related either to the m number of points included in the
415 subset I (for MDPTLT), or to the $\binom{n}{m}$ combinations of the subset I within the set T (for MDPT2T
416 and MDPTRT). Since the power of a statistical test is generally positively correlated to the sample
417 size, and provided that m and $\binom{n}{m}$ are quantities expressing the sample size, this finding is in
418 agreement with theory, despite this theory has been developed for a few tests and mostly in a
419 parametric framework (Acutis et al. 2012; Crawley 2013). Under a practical perspective, the
420 MDPTLT seems to be endowed with the best performances in terms of power, also when m and
421 $\binom{n}{m}$ are relatively small, while MDPT2T and MDPTRT appear to be more reliable when the ratio
422 m/n tends towards the 50%. The estimates of type I error do not seem to be a criterion allowing to
423 prefer one test to another according to the sampling size, as suggested by the almost complete lack
424 of correlation with the above mentioned parameters. Despite the MC simulations were performed
425 only for MDPT, they might be considered extendable to the corresponding MDRT, provided that B
426 is large enough to achieve reliable estimates of p_e . In fact, as stated before, the randomization tests
427 are unbiased approximations of their related permutation tests, whose accuracy can be improved up
428 to the desired level (Ernst 2004).

429 The assessment of power and type I error through MC simulations is a numerical validation,
430 since it is performed on known DGP. However, a biological validation is pivotal to verify the
431 performances of a statistical test in the field (Thébaud et al. 2005). The biological validation was
432 performed only on the MDRT in consideration of the above cited computational constraints.
433 However, the 95% confidence intervals of p_r indicate a good level of accuracy and exclude
434 ambiguity in the acceptance/rejection of the null hypotheses. Considering the combined results of
435 the three MDRT, the points displaying a $DR > 0$ within the network of sampling points covering the

436 study area were clustered for *H. annosum* and randomly distributed for *H. irregulare*. Thus, for both
437 fungal species, MDRT provided responses which were consistent among different tails and in
438 agreement with the results obtained by Gonthier et al. (2012) by using spatial autocorrelation
439 analyses, hence confirming the reliability of the MDRT in field conditions. Moreover, the
440 advantage of performing the MDRT rather than autocorrelation analysis is intrinsic in the
441 categorical measurement of the variable under investigation. The DR measured by Gonthier et al.
442 (2012) required the counting of all fungal colonies of *Heterobasidion* spp. under a dissecting
443 microscope, in addition to an appropriate sampling of colonies aimed at obtaining a large number of
444 isolates (up to 40 per sampling point). The molecular analyses performed on these isolates were the
445 last step to carry out the repartition of the DR between the two pathogenic species. This approach
446 provided a quantitative information, which was essential to compare spores deposition between the
447 two species as well as to carry out the autocorrelation analyses. However, the MDT could optimize
448 the experimental design in similar trials. In fact, the assessment of the condition $DR > 0$ could allow
449 a less refined sampling procedure. For instance, molecular analyses could be dramatically reduced
450 by pooling the samples of fungal mycelium of all isolates from each sampling point before DNA
451 extraction. Also the number of isolates could be probably reduced without a substantial loss of
452 information. Besides, the MDT could be performed on wide study areas, providing preliminary
453 results to be further investigated turning to the quantitative level, but only in representative
454 subareas.

455 The application of the MDT to the case study of the nut rot caused by *G. castanea* showed a
456 possible way through which the designed geostatistical tests can be performed to gather information
457 about a plant disease. Regardless of the area where the tests were performed, all MDT agreed in the
458 identification of a random spatial pattern of the chestnut trees displaying the presence of *G.*
459 *castanea* in at least one nut within the sampled trees. Since in half of the areas chestnuts were
460 clustered, while in the other half they were not, it could be argued that the plantation density is not a
461 variable influencing the spatial distribution of the pathogen. This conclusion seems to be confirmed

462 by the absence of significant differences among the incidences of the pathogen among the areas.
463 These findings suggest that the choice of the plantation density, which is a relevant issue for
464 chestnut growers (Dong-Sheng et al. 2009), can be based on other parameters (e.g. yield
465 productivity, intraspecific competition) rather than on the risk of transmission of *G. castanea*
466 among neighbouring trees. This finding is relevant since, to date, very little was known about the
467 relationship between the management practices and the incidence of *G. castanea*. However, it is
468 important to stress that results from geostatistics do not replace biological and epidemiological
469 investigations, but rather provide evidence about spatial distributions that can be helpful to
470 formulate and to test hypotheses about disease dynamics. In the case of *G. castanea* further analyses
471 are needed to determine the factors influencing the observed spatial patterns, since the infection
472 pathways of *G. castanea* are still mainly unknown (Lione et al. 2014).

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

482

483 **ACKNOWLEDGEMENTS**

484

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

487

488 **LITERATURE CITED**

489

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

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

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

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

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

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

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

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

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

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

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

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

- 513 2014. An integrated approach to control the introduced forest pathogen *Heterobasidion*
514 *irregulare* in Europe. *Forestry* 87:471-481.
- 515 Gonthier, P., Lione, G., Giordano, L., and Garbelotto, M. 2012. The American forest pathogen
516 *Heterobasidion irregulare* colonizes unexpected habitats after its introduction in Italy. *Ecol.*
517 *Appl.* 22:2135-2143.
- 518 Goslee, S.C. and Urban, D.L. 2007. The ecodist package for dissimilarity-based analysis of
519 ecological data. *J. Stat. Sofw.* 22:1-19.
- 520 Gottwald, T.R., Avinent, L., Llácer, G., de Mendoza, A.H., and Cambra, M. 1995. Analysis of the
521 spatial spread of sharka (plum pox virus) in apricot and peach orchards in eastern Spain. *Plant*
522 *Dis.* 79:266-278.
- 523 Gray, S.M., Moyer, J.W., and Bloomfield, P. 1986. Two-dimensional distance class model for
524 quantitative description of virus-infected plant distribution lattices. *Phytopathology* 76:243-
525 248.
- 526 Hartwig, F.P. 2013. Two-tailed p-values calculation in permutation-based tests: a warning against
527 “asymptotic bias” in randomized clinical trials. *J. Clin. Trials* 3: 145.
- 528 Jaime-Garcia, R., Trinidad-Correa, R., Felix-Gastelum, R., Orum, T.V., Wasmann, C.C., and
529 Nelson, M.R. 2000. Temporal and spatial patterns of genetic structure of *Phytophthora*
530 *infestans* from tomato and potato in the Del Fuerte Valley. *Phytopathology* 90:1188-1195.
- 531 Lamichhane, J.R., Fabi, A., Ridolfi, R., and Varvaro, L. 2013. Epidemiological study of hazelnut
532 bacterial blight in central Italy by using laboratory analysis and geostatistics. *PLoS ONE*, 8(2),
533 e56298. doi:10.1371/journal.pone.0056298.
- 534 Lione, G., Giordano, L., Sillo, F. and Gonthier, P. 2014. Testing and modelling the effects of
535 climate on the incidence of the emergent nut rot agent of chestnut *Gnomoniopsis castanea*.
536 *Plant Pathol.* doi: 10.1111/ppa.12319.
- 537 Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer*
538 *Res.* 27:209-220.

- 539 Marchant, B.P., and Lark, R.M. 2004. Estimating variogram uncertainty. *Math. Geol.* 36:867-898.
- 540 Mitchell, A. 2009. *The ESRI Guide To GIS Analysis. Volume 2. Spatial Measurements and*
541 *Statistics.* Environmental Systems Research Institute Press, Redlands, California, USA.
- 542 Nelson, S.C. 1995. STCLASS - spatiotemporal distance class analysis software for the personal
543 computer. *Plant Dis.* 79:643-648.
- 544 Nelson, S.C., Marsh, P., and Campbell, C. 1992. 2DCLASS, a two-dimensional distance class
545 analysis software for the personal computer. *Plant Dis.* 76:427-432.
- 546 Nelson, M.R., Orum T.V., Jaime-Garcia R., and Nadeem, A. 1999. Applications of geographic
547 information systems and geostatistics in plant disease epidemiology and management. *Plant*
548 *Dis.* 83:308-319.
- 549 Peres-Neto, P.R., and Olden, J.D. 2001. Assessing the robustness of randomization tests: examples
550 from behavioural studies. *Anim. Behav.* 61:79-86.
- 551 Pizzato, J.A., Araújo, D.V., Galvanin, E.A., Júnior, J.R., Matos, Â.N., Vecchi, M., and Zavislak,
552 F.D. 2014. Geostatistics as a methodology for studying the spatiotemporal dynamics of
553 *Ramularia areola* in cotton crops. *Am. J. Plant Sci.* 5:2472-2479.
- 554 Thébaud, G., Peyrard, N., Dallot, S., Calonnec, A., and Labonne, G. 2005. Investigating disease
555 spread between two assessment dates with permutation tests on a lattice. *Phytopathology*
556 95:1453-1461.
- 557 Visentin, I., Gentile, S., Valentino, D., Gonthier, P., Tamietti, G., and Cardinale F. 2012.
558 *Gnomoniopsis castanea* sp. nov (Gnomoniaceae, Diaporthales) as the causal agent of nut rot
559 in sweet chestnut. *J. P. P.* 94:411-419.
- 560 Webster, R., and Oliver, M.A. 2001. *Geostatistics For Environmental Scientists.* John Wiley and
561 Sons Ltd., Chichester, UK.
- 562 Weidong, L. 2006. Transiogram: a spatial relationship measure for categorical data. *Int. J.*
563 *Geograph. Inf. Sci.* 20:693-699.
- 564

565 **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	- \bar{d}_0 : observed mean value of the triangular Euclidean distance matrix among the points in I
	MDPTLT	left-tailed	the spatial pattern of level γ is not clustered	- α : significance level cut-off	- \bar{D} : mean of the permutation distribution
	MDPTRT	right-tailed	the spatial pattern of level γ is not dispersed		- p_e : exact p-value
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	\bar{d}_0 : observed mean value of the triangular Euclidean distance matrix among the points in I
	MDRTLTL	left-tailed	the spatial pattern of level γ is not clustered	- α : significance level cut-off - B : number of random combinations	- \bar{D} : mean of the randomization distribution
	MDRTRT	right-tailed	the spatial pattern of level γ is not dispersed	- λ : confidence level for the p-value	- 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

566

567

568

569

Table 2

Test	DGP verifying H₀	DGP not verifying H₀	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

570

571

572

573

574

575

576

577

578

579

580

581

582 **Table 3.**

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

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

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

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

589

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

593

594 Table 3. Estimates of power and type I error for the Mean Distance Permutation Tests (MDPT)
 595 obtained through Monte Carlo simulations and results of the correlation analysis. The estimates are
 596 provided for each block of simulations ranked according to the m values and divided for two-tailed,
 597 left-tailed and right-tailed tests (MDPT2T, MDPTRT, and MDPTLT) and window size. The

598 number of combinations $\binom{n}{m}$ enumerated for each value of m is listed. The average of power and

599 type I error as well as the Spearman correlation coefficient between the estimates and m [i.e. $\rho(m)$]

600 and $\binom{n}{m}$ [i.e. $\rho\left(\binom{n}{m}\right)$] are reported with the related p-value for all tests and window sizes. The

601 symbol * indicates correlation coefficients significant at 0.05 cut-off.

602

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

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

610

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

614

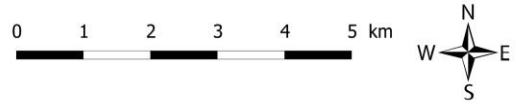
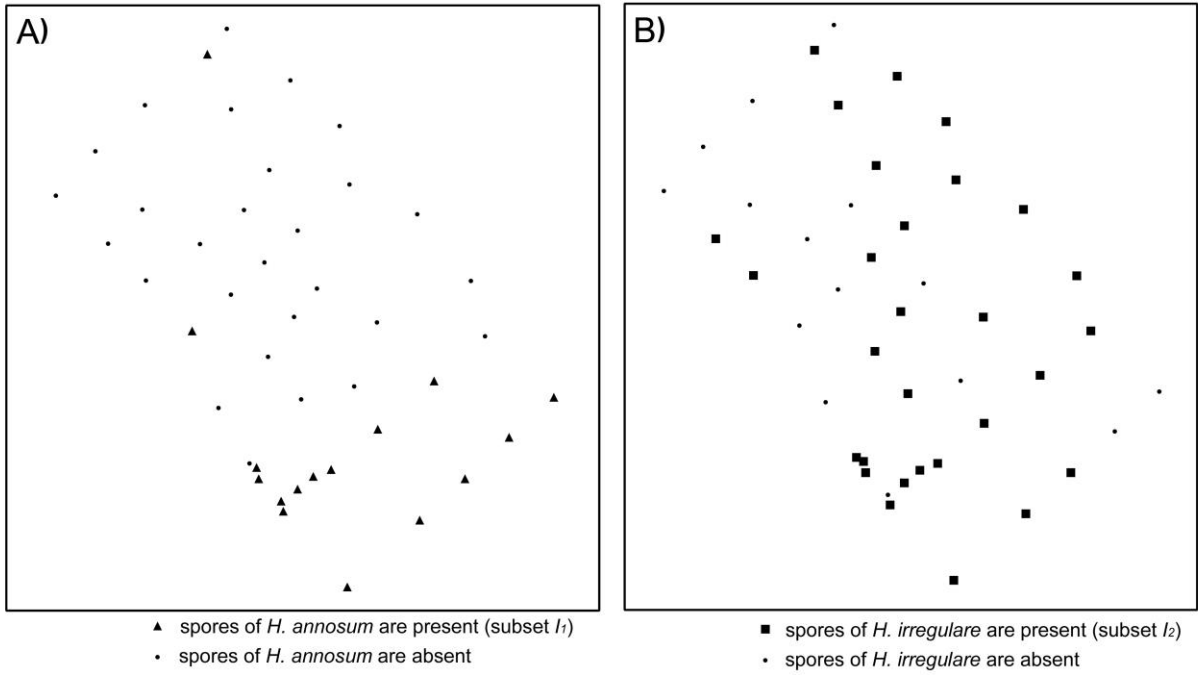
615 Fig. 2. Maps of chestnut trees of the “Vivaio Gambarello” orchard carrying at least one nut infected
616 by *Gnomoniopsis castanea* (level γ) in areas C1 (A), C2 (B), NC1 (C) and NC2 (D).

617

618

619 FIG. 1

620

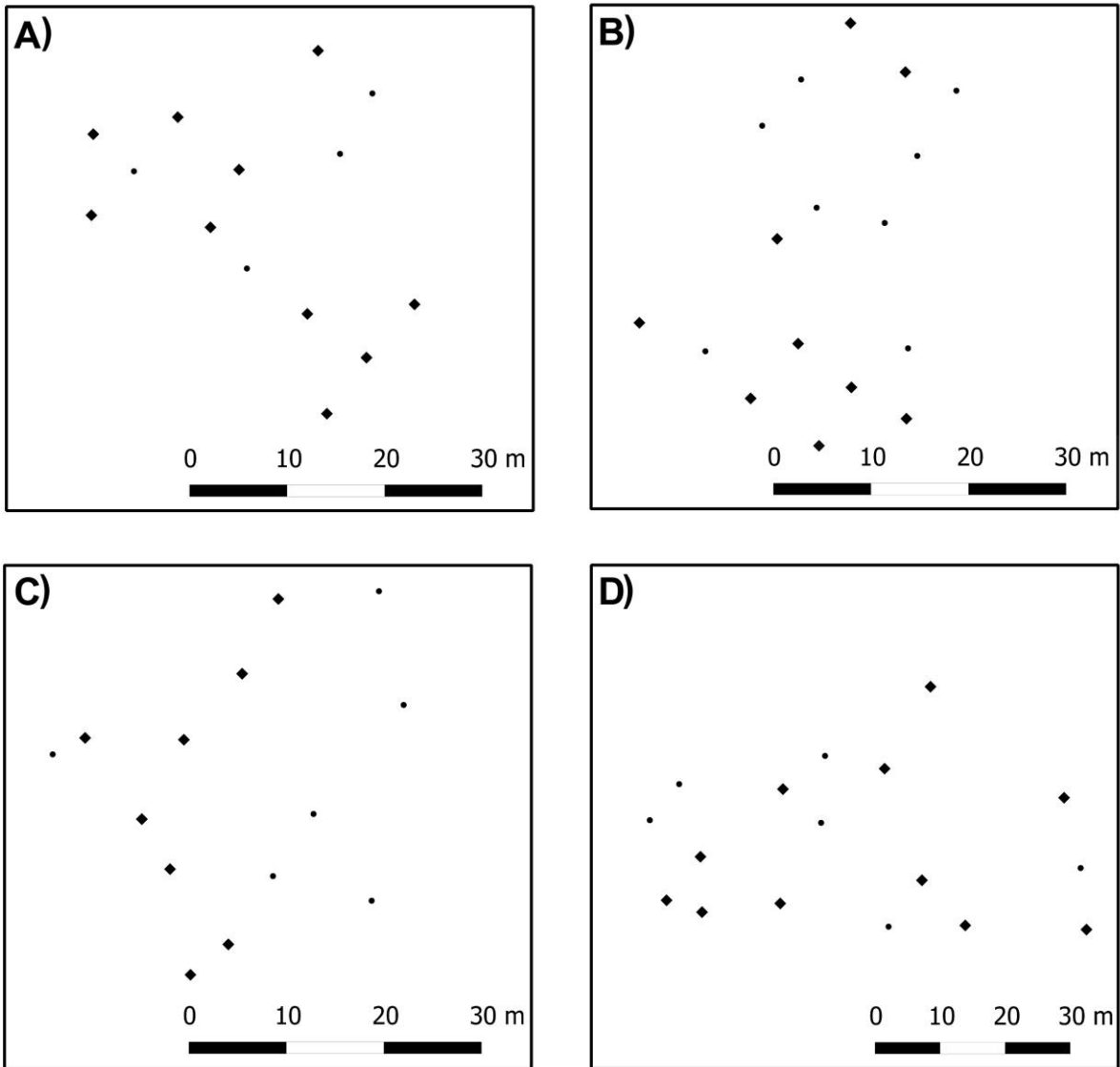


621

622

623

624 FIG. 2



- ◆ *G. castanea* is present in at least one nut (level γ)
- *G. castanea* is not present in the nuts



625

626

627