Original Research
Vadose Zone Journal

Site Classification to Predict the Abundance of the Deep-Burrowing Earthworm *Lumbricus terrestris* L.

A. M. L. Lindahl,* I. G. Dubus, and N. J. Jarvis

Channels made by deep-burrowing (anecic) earthworms are known to strongly affect soil water flow and increase the leaching risk of agricultural pollutants. A classification tree that predicts the abundance of the anecic earthworm *Lumbricus terrestris* L. from readily available survey information (land use, management practices, and soil texture) was derived from literature data (n = 86). The most important factors favoring *L. terrestris* were perennial land use, no-till arable cropping, organic additions (i.e., manure), and medium-textured soil. The classification scheme correctly predicted earthworm abundance for 71% of the studies in the database. Among other potential applications, the classification tree could be used to identify areas at risk from groundwater pollution in agricultural landscapes and to support catchment- and regional-scale models of contaminant leaching in the vadose zone.

Earthworms have been termed "ecosystem engineers" (Jones et al., 1994) since they significantly influence many important soil processes by modifying the structure and architecture of the soil. Among other processes, earthworm activities strongly affect soil hydrology. For example, under favorable conditions, populations of surface-feeding anecic earthworm species create permanent burrows, penetrating deep into the subsoil, that have been shown to significantly increase ponded infiltration rates and saturated hydraulic conductivity (e.g., Bouché and Al-Addan, 1997; Shipitalo et al., 2004). Under natural field conditions, water infiltrating in these large earthworm macropores rapidly and preferentially bypasses the soil unsaturated zone (e.g., Ehlers, 1975; Germann et al., 1984; Edwards et al., 1992), which has major implications for the risk of pollution of groundwater and surface water (Jarvis, 2007). In contrast, the effects of endogeic earthworms on water flow and solute transport may be more limited (Ela et al., 1991), since they primarily feed within the topsoil and produce temporary burrows that are more randomly oriented, tortuous, and branched (Capowiez et al., 2001; Jégou et al., 2001).

Populations of anecic earthworms can be spatially highly variable. Under favorable conditions, populations can amount to several hundred individuals per square meter, whereas they may be completely absent if one or more environmental factors is unfavorable (Lee, 1985). This suggests that an improved quantitative understanding of the site and soil factors controlling the abundance of anecic earthworms in agroecosystems could help to improve methods to identify areas at risk from groundwater pollution. In principle, the degree of preferential flow in soil is dependent on the density of conducting earthworm burrows (Smettem, 1992) rather than population size. The density of water-conducting burrows, however, has been shown to be correlated with the size of anecic earthworm populations (Chan and Heenan, 1993; Bouché and Al-Addan, 1997; Chan, 2004), which is useful since there are far more data in the literature concerning the latter than the former.

Many previous studies have provided useful insights into site factors that control earthworm abundance. Anecic earthworm populations are, for example, favored by perennial crops and conservation tillage systems in arable agriculture, which improve the food supply of fresh litter at the soil surface (Ehlers, 1975; Barnes and Ellis, 1979; Edwards and Lofty, 1982; Lee, 1985; Edwards et al., 1992; Chan, 2001; Jordan et al., 2004). Fundamental soil properties such as pH, organic matter content, texture, and moisture status also influence earthworm population densities (e.g., Lee, 1985; Hendrix et al., 1992; Poier and Richter, 1992;

A.M.L. Lindahl and N.J. Jarvis, Dep. of Soil and Environment, Swedish Univ. of Agricultural Sciences, 750 07 Uppsala, Sweden; I.G. Dubus, BRGM/FOOTPRINT, Ave. C. Guillemin, BP36009, 45060 Orléans Cedex 2, France. *Corresponding author (anna.lindahl@mark.slu.se).

Vadose Zone J. 8:911–915
doi:10.2136/vzj2008.140
Received 3 Oct. 2008.
Published online 13 Sept. 2009.

Cannavacciuolo et al., 1998; Joschko et al., 2006; Ouellet et al., 2008). Most previous studies have been performed at the plot or field scale, which necessarily restricts the number of environmental factors that can be studied. The few published regional-scale studies (e.g., Joschko et al., 2006; Ouellet et al., 2008) also suffer from a similar disadvantage in that, for practical reasons, they covered only a limited number of sites within a restricted range of agroenvironmental conditions. In this study, we developed a quantitative scheme to predict the abundance of the anecic earthworm *Lumbricus terrestris* in agroecosystems from easily obtained site and soil factors. The scheme is based on a statistical data mining analysis of literature data using the method of classification trees. *Lumbricus terrestris* was selected as a model for anecic earthworms, since it is widely distributed and is probably the most studied of all earthworm species. In a companion study, Jarvis et al. (2009) described how the scheme is being used to support predictions of site susceptibility to macropore flow within a broader risk assessment framework (FOOTPRINT, www.eu-footprint.org; verified 30 Mar. 2009) for pesticide leaching in the vadose zone. Although this was the primary motivation for our study, the classification scheme derived in this study may have other potential applications, considering the role of earthworms as "ecosystem engineers."

Materials and Methods

Literature Study

A database on the abundance of *L. terrestris* in agricultural soil was created from literature data published between 1961 and 2006. A total of 39 scientific studies covering 86 agricultural sites across Europe were identified. Studies from other parts of the world where *L. terrestris* has been introduced and spread by humans were deliberately excluded. In the studies included in the database, the total number of *L. terrestris* ranged from 0 to 220 individuals m^{-2}. The methods used for estimating the number of worms included application of chemicals (n = 54), electrical impulses (n = 4), flooding (n = 1), digging and hand sorting (n = 12), or a combination of methods (n = 15). In six studies, only the number of adult earthworms was reported. For these studies, the total number was estimated from the ratio of the number of adult earthworms per juvenile. This reference ratio (=0.3, SE = 0.045) was estimated by linear regression from the data in 18 studies that reported numbers of both juvenile and adult earthworms (R^2 = 0.5, $P < 8 \times 10^{-6}$). In those few cases where no specific information was given, we assumed that the number of worms reported represented the total population (i.e., both adults and juveniles). For those studies in which populations were measured on multiple occasions in any 1 yr, data were only taken from seasons when the earthworms could be expected to be most active (i.e., spring). Averages were calculated when data from multiple years were available.

The resulting database comprises study-specific information (reference to the study, year of publication, date of execution, method of worm extraction, and sampling depth) and the reported number of worms per square meter. The database records also include information on the location of the study site, its location defined according to the 16 European climatic zones (Blenkinsop et al., 2008; Centofanti et al., 2008) derived as part of the FOOTPRINT project (www.eu-footprint.org), soil depth and site hydrologic conditions, as well as land use and management practices (cropping, tillage, and manuring) and soil properties (pH, organic C content, and texture). The database is not complete since no study has reported all of the above-mentioned information. It is interesting to note that many studies have provided little information concerning the properties and characteristics of the habitat of the earthworm, the soil. The database can be obtained from A.M.L. Lindahl.

Classification Trees

Classification tree data mining techniques were used to develop a systematic and quantitative method to predict the abundance of *L. terrestris* in agricultural soil. The resulting classification scheme had to fulfill three requirements. It should (i) yield reliable predictions, (ii) provide insight into how earthworm abundance relates to soil and site factors, and (iii) be compatible with a broader decision-support system for predictive modeling of pesticide leaching (Jarvis et al., 2009). Classification trees fit our purposes well since they produce visual tree structures that are easily understood and interpreted (Witten and Frank, 2005).

Tree-structured classifiers are constructed by selecting the most useful variables from a set of candidate predictor variables to sequentially split the learning data into descendant subsets of purer class membership (Breiman et al., 1984). The construction revolves around selecting splits, deciding whether to continue splitting each subset or to stop splitting by creating a terminal node, and assigning each terminal node to a class. Tree size is constrained by statistics that maximize the predictive capacity while minimizing the tree size. We used the C4.5 algorithm, a statistical classifier developed by Quinlan (1993), to generate the decision tree. The C4.5 decision tree is a simple univariate tree, such that only one attribute is used to split the data into subsets at each node of the tree. The attribute with the highest normalized information gain (difference in entropy) is chosen to split the data. The subsets are recursively split until the user-defined tree-growing stopping limits are reached (i.e., the minimum number of instances permissible at a terminal node). Finally, the created tree is pruned by removing branches that do not improve the performance of the tree and replacing them with terminal nodes. The measure of predictive power is the error rate, which is a measure of the overall performance of the classifier. The error rate is calculated as the proportion of errors made across a whole set of instances. The upper confidence limit is set as a pessimistic estimate of the error rate. A tree node is pruned if the error estimate at the node is less than the combined error estimate of the descendant nodes.

Since error rates on training sets are optimistic, the error rate must be calculated using test sets, but the amount of data for training and testing are often limited. A common way to solve this issue is to use stratified 10-fold cross-validation (Witten and Frank, 2005). This technique first divides the data randomly into 10 parts, trains the learning scheme on nine of the subsets, and then calculates the error rate on the 10th set. This procedure is executed 10 times so that each of the 10 subsets is used once for calculating the error rate. The overall error estimate is finally calculated by averaging the 10 error estimates. One single 10-fold cross-validation might not be sufficient to obtain a realistic error estimate. It is therefore standard procedure to repeat the process 10 times and average the 100 resulting error estimates.

We used the WEKA software (Witten and Frank, 2005) Version 3.5.6 to apply the C4.5 algorithm to our database on the abundance of *L. terrestris*, using default values of two for the minimum number of instances as a tree-growing stopping limit and a confidence limit of 25% for pruning. Stratified 10-fold cross-validation was repeated 10 times to estimate the reliability of the classification tree.

Eight of the variables in the database were selected as potential predictors (see Table 1). The abundance of *L. terrestris* at each site in the database was a priori and arbitrarily classed as low (<3 m^{-2}), medium (3–10 m^{-2}), or high (>10 m^{-2}). These cutoff values were selected to achieve an approximately equal distribution of sites between the classes (27 low abundance sites, 30 sites of medium abundance, and 29 of high abundance).

Results and Discussion

A classification tree was derived (Fig. 1) that is easy to interpret and correctly identifies earthworm abundance for 71% of the 86 studies in the database. The classification tree correctly predicted 86% of the high-abundance sites and 63% of the medium- and low-abundance sites. Pure guessing, based on the number of instances of each abundance class in the database, would give success rates of 31 to 35%. On the basis of the 10, 10-fold cross-validations, the expected accuracy of the classification tree applied to other agricultural sites in Europe is 65%.

The abundance of *L. terrestris* in agroecosystems was found to be strongly influenced by land use, management practices, and soil texture. The tree originally comprised four predictor variables, but became a simple two-variable tree by combining the three predictor variables of land use, tillage system, and organic additions. The variables climate, depth restriction, hydrologic conditions, and pH were not included, since they did not improve the predictive power of the classification. The C4.5 algorithm probably did not select either pH or depth restriction as useful variables on which to split because very few studies had either a shallow soil depth or a pH <5 (see Table 1). Thus, one reason why it may be difficult to identify some potentially important limiting environmental conditions by data mining analysis of field experiments is that researchers rarely choose to investigate earthworm abundance at sites where they are not present. This lack of data may have been exacerbated by correlations and interactions between predictor variables. For example, *L. terrestris* was absent at two arable study sites in the database where shallow rock occurred at depths <25 cm (Barnes and Ellis, 1979), but were found in abundance on a similar shallow soil type (chalky rendzina) at an undisturbed

Table 1. Predictor variables tested in the classification tree analysis.

Variable	No. of studies
Climate	
Cold (=FCZ 4,6,~~10~~)†	25
Other (=FCZ 1,2,3,~~5~~,7,~~8~~, ~~9~~,11,~~12~~,13,14,15,16)	61
Depth restriction	
Yes (soil depth ≤25 cm)	3
No	83
Hydrologic condition	
Free draining	31
Stagnogleyic (slowly permeable)	36
High water table	1
Land use	
Arable	54
Ley or mixed rotation	10
Perennials (grassland or orchards)	22
Texture (USDA)	
Fine (silty clay, clay, clay loam, or silty clay loam)	26
Medium (other texture classes)	55
Coarse (loamy sand or sand)	5
Tillage system	
Conventional	31
Reduced	9
No-till	14
Organic additions	
Yes	9
No	77
pH	
<5	3
>5	42

† Footprint climate zones (see Centofanti et al., 2008); strike-through parameter values are not represented in the database.

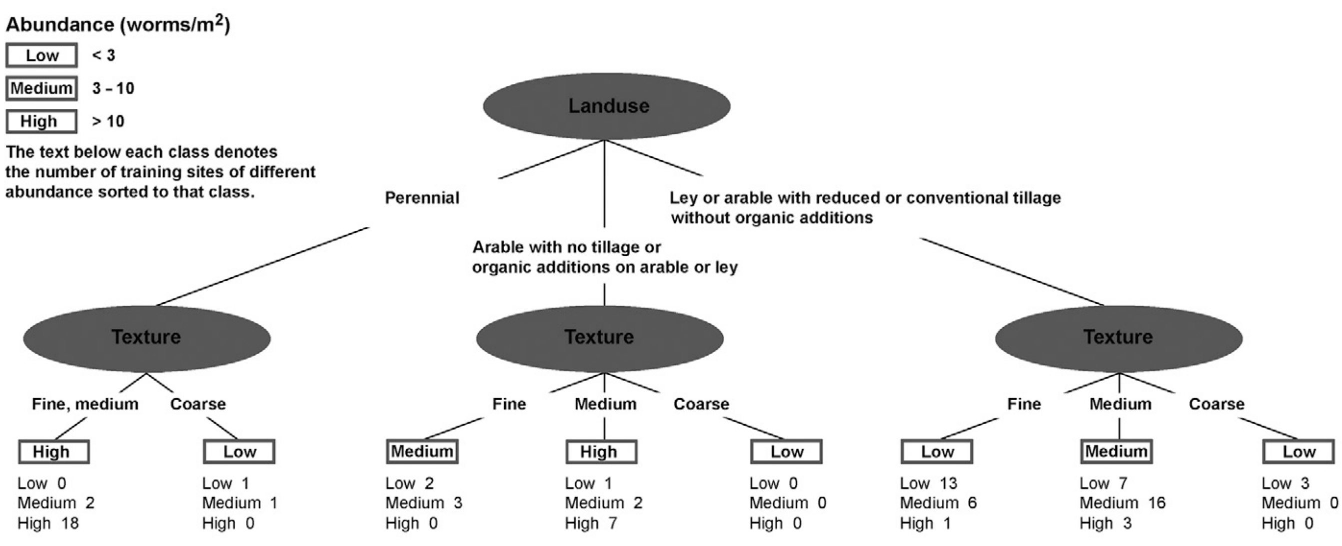

Fig. 1. Classification tree for predicting the abundance of *Lumbricus terrestris* based on data from 86 agricultural sites in Europe.

grassland site (Margerie et al., 2001). All three sites with low pH also had sandy-textured soil, which also seems to restrict earthworm abundance (see below). This correlation would prevent the algorithm from identifying low pH as a limiting factor.

Land use was the most important predictor variable (i.e., having the highest normalized information gain). The classification tree confirmed that perennial land uses (grassland and orchards) are very beneficial to anecic earthworms like *L. terrestris* (Fig. 1), presumably due to the lack of disturbance and also because they improve the supply of food (fresh organic material) compared with arable cropping systems (Lee, 1985). According to the classification tree, the least favorable land use systems are ley or arable with reduced or conventional tillage without organic additions. Plow-based tillage systems are unfavorable for anecic earthworms because they disrupt their burrows, increase predation, and reduce the food supply (i.e., aboveground crop residues). There were no terminal nodes assigned as high abundance for these land uses and management practices. Compared with conventional arable systems, the food supply to earthworms is improved if either manure is applied or direct drilling (zero-till) is practiced. The classification tree confirmed that such systems are intermediate between perennial and conventional arable land uses.

Texture was selected as the second most important predictor variable. Medium-textured soils are clearly most favorable for anecic earthworms (Fig. 1). Coarse-textured soil appears to be the least favorable, with four out of five instances in the database falling in the low-abundance class (Fig. 1). Sandy soils are thought to be abrasive to burrowing earthworms and are also drought prone (Lee, 1985). A terminal node assigned to low abundance for coarse-textured soil was added to the tree for the land use category "arable with no tillage or organic additions on arable or ley," even though the database does not contain any such instances (Fig. 1). The classification tree also shows that fewer earthworms are found in clayey soils than in medium-textured soils. This may be due to several different factors, including a higher resistance to burrowing and a greater risk of waterlogging (Lee, 1985).

Conclusions

The abundance of the anecic earthworm *L. terrestris* in agroecosystems can be predicted by the classification tree developed in this study from four site and soil factors: land use, tillage system, organic additions, and soil texture. The tree is very easy to use and interpret due to its small size and the fact that the predictor variables are easily obtained from survey information. These are attractive characteristics for a quantitative scheme that is to be used in a decision-support system for regional-scale predictive modeling of pesticide leaching.

Acknowledgments

This research was undertaken as part of the FORMAS (Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning) funded project "Understanding the variability of pesticide losses to surface waters and groundwater at field and catchment scales" and the EU project FOOTPRINT (Functional tools for pesticide risk assessment and management, Project no. SSPI-CT-2005-022704). Funding by the European Commission through its 6th Framework Program for Research and Development is gratefully acknowledged.

References

Barnes, B.T., and F.B. Ellis. 1979. Effects of different methods of cultivation and direct drilling, and disposal of straw residues, on populations of earthworms. J. Soil Sci. 30:669–679.

Blenkinsop, S., H.J. Fowler, I.G. Dubus, B.T. Nolan, and J.M. Hollis. 2008. Developing climatic scenarios for pesticide fate modelling in Europe. Environ. Pollut. 154:219–231.

Bouché, M.B., and F. Al-Addan. 1997. Earthworms, water infiltration and soil stability: Some new assessments. Soil Biol. Biochem. 29:441–452.

Breiman, L., J.H. Freidman, R.A. Olshen, and C.J. Stone. 1984. Classification and regression trees. Wadsworth, Monterey, CA.

Cannavacciuolo, M., A. Bellido, D. Cluzeau, C. Gascuel, and P. Trehen. 1998. A geostatistical approach to the study of earthworm distribution in grassland. Appl. Soil Ecol. 9:345–349.

Capowiez, Y., P. Monestiez, and L. Belzunces. 2001. Burrow systems made by *Aporrectodea nocturna* and *Allolobophora chlorotica* in artificial cores: Morphological differences and effects of interspecific interactions. Appl. Soil Ecol. 16:109–120.

Centofanti, T., J.M. Hollis, S. Blenkinsop, H.J. Fowler, I. Truckell, I.G. Dubus, and S. Reichenberger. 2008. Development of agro-environmental scenarios to support pesticide risk assessment in Europe. Sci. Total Environ. 407:574–588.

Chan, K.Y. 2001. An overview of some tillage impacts on earthworm population abundance and diversity: Implications for functioning in soils. Soil Tillage Res. 57:179–191.

Chan, K.Y. 2004. Impact of tillage practices and burrows of a native Australian anecic earthworm on soil hydrology. Appl. Soil Ecol. 27:89–96.

Chan, K.Y., and D.P. Heenan. 1993. Surface hydraulic properties of a red earth under continuous cropping with different management practices. Aust. J. Soil Res. 31:13–24.

Edwards, C.A., and J.R. Lofty. 1982. The effect of direct drilling and minimal cultivation on earthworm populations. J. Appl. Ecol. 19:723–734.

Edwards, W.M., M.J. Shipitalo, S.J. Traina, C.A. Edwards, and L.B. Owens. 1992. Role of *Lumbricus terrestris* (L.) burrows on quality of infiltrating water. Soil Biol. Biochem. 24:1555–1561.

Ehlers, W. 1975. Observations on earthworm channels and infiltration on tilled and untilled loess soil. Soil Sci. 119:242–249.

Ela, S.D., S.C. Gupta, W.J. Rawls, and J.F. Moncrief. 1991. Role of earthworm macropores formed by *Aporrectodea tuberculata* on preferential flow of water through a Typic Hapludoll. p. 68–76. *In* T.J. Gish and A. Shirmohammadi (ed.) Preferential flow. Proc. Natl. Symp., Chicago. 16–17 Dec. 1991. ASAE, St. Joseph, MI.

Germann, P.F., W.M. Edwards, and L.B. Owens. 1984. Profiles of bromide and increased soil moisture after infiltration into soils with macropores. Soil Sci. Soc. Am. J. 48:237–244.

Hendrix, P.F., B.R. Mueller, R.R. Bruce, G.W. Langdale, and R.W. Parmelee. 1992. Abundance and distribution of earthworms in relation to landscape factors on the Georgia Piedmont, U.S.A. Soil Biol. Biochem. 24:1357–1361.

Jarvis, N.J. 2007. A review of non-equilibrium water flow and solute transport in soil macropores: Principles, controlling factors and consequences for water quality. Eur. J. Soil Sci. 58:523–546.

Jarvis, N.J., J. Moeys, J.M. Hollis, S. Reichenberger, A.M.L. Lindahl, and I.G. Dubus. 2009. A conceptual model of soil susceptibility to macropore flow. Vadose Zone J. 8:902–910 (this issue).

Jégou, D., Y. Capowiez, and D. Cluzeau. 2001. Interactions between earthworm species in artificial soil cores assessed through the 3D reconstruction of the burrow systems. Geoderma 102:123–137.

Jones, C.G., J.H. Lawton, and M. Shachak. 1994. Organisms as ecosystem engineers. Oikos 69:373–386.

Jordan, D., R.J. Miles, V.C. Hubbard, and T. Lorenz. 2004. Effect of management practices and cropping systems on earthworm abundance and microbial activity in Sanborn Field: A 115-year-old agricultural field. Pedobiologia 48:99–110.

Joschko, M., C.A. Fox, P. Lentzsch, J. Kiesel, W. Hierold, S. Krück, and J. Timmer. 2006. Spatial analysis of earthworm biodiversity at the regional scale. Agric. Ecosyst. Environ. 112:367–380.

Lee, K.E. 1985. Earthworms: Their ecology and relationship with soils and land use. Academic Press, Sydney, NSW, Australia.

Margerie, P., T. Decaëns, F. Bureau, and D. Alard. 2001. Spatial distribution of earthworm species assemblages in a chalky slope of the Seine Valley (Normandy, France). Eur. J. Soil Biol. 37:291–296.

Ouellet, G., D.R. Lapen, E. Topp, M. Sawada, and M. Edwards. 2008. A heuristic model to predict earthworm biomass in agroecosystems based on selected management and soil properties. Appl. Soil Ecol. 39:35–45.

Poier, K.R., and J. Richter. 1992. Spatial distribution of earthworms and soil properties in an arable loess soil. Soil Biol. Biochem. 24:1601–1608.

Quinlan, J.R. 1993. C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco.

Shipitalo, M.J., V. Nuutinen, and K.R. Butt. 2004. Interaction of earthworm burrows and cracks in a clayey, subsurface-drained soil. Appl. Soil Ecol. 26:209–217.

Smettem, K.R.J. 1992. The relation of earthworms to soil hydraulic properties. Soil Biol. Biochem. 24:1539–1543.

Witten, I.H., and K. Frank. 2005. Data mining: Practical machine learning tools and techniques. 2nd ed. Morgan Kaufmann, San Francisco.