ΦYTON
REVISTA INTERNACIONAL DE BOTÁNICA EXPERIMENTAL
INTERNATIONAL JOURNAL OF EXPERIMENTAL BOTANY
FUNDACION ROMULO RAGGIO
Gaspar Campos 861, 1638 Vicente López (BA), Argentina
www.revistaphyton.fund-romuloraggio.org.ar

Application of the random forest algorithm for predicting the persistence of seed banks in the Horqin Sandy Land, China

Aplicación del algoritmo del bosque al azar para predecir la persistencia de los bancos de semilla en el área arenosa de Horquin, China

Tang Y[1] & SS Jin[2]

Abstract. Persistent seed banks have been detected in the Horqin Sandy Land, China using experimental methods. In this study, we used seed traits (i.e. seed mass and seed shape) to predict the persistence of seed banks using the random forest algorithm. The results showed that the mean decrease in accuracy for seed mass and seed shape was 18.26 and 9.90, respectively, suggesting that seed mass was a better predictor than seed shape. With increasing seed mass, the log of P (where P is the ratio of the number of votes selecting existence of a persistent seed bank to the number of votes selecting absence of a persistent seed bank) gradually decreased. It also changed from positive to negative when the seed mass reached 10 mg. This indicates that small-seeded species tend to have a persistent seed bank, and larger-seeded species tend to germinate in the current year. Furthermore, a seed mass of 10 mg was the dividing point for predicting the persistence of seed banks in this region. The log of P decreased as the seed shape decreased in the range from 0.07 to 0.17, indicating that seed shape (0.07–0.17) is negatively related to the existence of a persistent seed bank. It is suggested that the random forest algorithm is a useful tool for predicting the persistence of seed banks.

Keywords: Model; R software; Sandy lands; Seed mass; Seed shape.

Resumen. Se han detectado bancos de semillas persistentes en el área arenosa de Horqin, China, usando métodos experimentales. En este estudio, usamos características de las semillas (por ej., peso y forma de la semilla) para predecir la persistencia de los bancos de semilla usando el algoritmo del bosque al azar. Los resultados mostraron que la reducción promedio en exactitud para el peso y la forma de la semilla fueron 18,26 y 9,90, respectivamente, sugiriendo que el peso de la semilla fue un mejor predictor que la forma de la semilla. Con incrementos en el peso de la semilla, el logaritmo de P disminuyó gradualmente (donde P es la relación del número de votos que seleccionan por la existencia de un banco de semillas persistente al número de votos que seleccionan por la ausencia de un banco de semillas persistentes). También cambió de positivo a negativo cuando el peso de la semilla alcanzó los 10 mg. Esto indica que las especies que producen semillas pequeñas tienden a tener un banco de semillas persistente, y que las especies que producen semillas más pesadas tienden a germinar en el año en que fueron producidas. Además, un peso de semillas de 10 mg fue el punto de división para predecir la persistencia de bancos de semillas en la región de estudio. El logaritmo de P disminuyó cuando la forma de la semilla disminuyó en el rango desde 0,07 a 0,17, indicando que la forma de la semilla (0,07-0,17) está negativamente relacionada a la existencia de un banco de semillas persistentes. Se sugiere que el algoritmo del bosque al azar es una herramienta útil para predecir la persistencia de bancos de semillas.

Palabras clave: Modelo; Software R; Zonas con suelos arenosos; Peso de la semilla; Forma de la semilla.

[1] School of Life Science, Liaoning University, Shenyang, 110036, China.
[2] School of Management, Fudan University, Shanghai 200433, China.
Address correspondence to: Shusong Jin, *e-mail:* Jinss@fudan.edu.cn
Received 22.X.2016. Accepted 8.VIII.2017.

INTRODUCTION

Seed banks in soils provide opportunities for plants to escape from disadvantageous habitats, especially in arid and semi-arid lands where plants are exposed to droughts and disturbances, such as grazing (Ariki & Washitani, 2000; Yan, et al. 2005; Saatkamp et al., 2009; Tang et al., 2014). Soil seed banks play an important role in understanding the life history of plant populations and exploring the relationship between plants and environments (Ma & Liu, 2008; Borgy et al., 2015). Persistent seed banks, defined as seeds that persist in the soil for more than 1 year, are included in seed bank classification systems (Thompson, 1993a; Yu et al., 2007). A persistent seed bank is a reproductive strategy for special plant species, and is helpful for ensuring the potential for plant population growth in disadvantageous habitats and maintaining vegetation after disturbances (Yan et al., 2005a; Liu et al., 2007).

Commonly, persistent seed banks are detected using experimental methods. Saatkamp et al. (2009) listed seven methods to study the persistence of seeds in the soil seed bank, including seed burial experiments. However, these experimental methods require considerable time and labor. Recently, a number of studies have focused on exploring the relationship between the persistence of seeds and seeds traits (Thompson et al., 1993b; Zhao et al., 2011; Schutte et al., 2014), seeking that the persistence of seeds can be predicted through these seeds traits. For example, seed size and seed shape have been found to be correlated with the persistence of seeds in sandy lands (Chao et al., 2011). The persistence of seeds could be explained by seed size and seed coat thickness (Schutte et al., 2014). However, how to use seed traits to predict seed persistence in soils is not clearly known.

The existence and absence of a persistent seed bank can be described by a binary variable (0/1), a type of categorical variable. Therefore, prediction of the existence of persistent seed banks is a classification issue, which could be solved using the random forest algorithm. The random forest algorithm, developed in 2001, consists of decision trees, each of which gives a classification. The forest chooses the classification with the most votes (Breiman, 2001). Recently, the random forest algorithm has been widely used by ecologists and botanists to predict species distributions, to improve the accuracy of spatial interpolation of environmental variables, and to explore the relationship between environmental variables and family ages of special communities (Li et al., 2011; Liu et al., 2013; Crego et al., 2014). Overall, the random forest algorithm has proved to be one of the most accurate learning algorithms dealing with classification issues (Li & Wang, 2013).

In the Horqin Sandy Land, China, where sand burial and wind erosion are unfavorable for the recruitment and survival of plant species, a persistent seed bank is helpful for the survival of plant species in this unstable habitat (Ma & Liu, 2008). For instance, a persistent seed bank is critical for *Agriophyllum squarrosum*, a pioneer annual plant species, to adapt to sand mobility (Liu et al., 2006). Thus, persistent seed banks of plant species in the Horqin Sandy Land have received much attention. We hope to predict the existence of persistent seed banks using seeds traits. Moreover, we want to determine the relative importance of seed traits when predicting the existence of persistent seed banks.

The aims of this study were to: (1) test whether the random forest algorithm method is suitable for predicting the existence of persistent soil seed banks; (2) explore the relationship between seeds traits and the persistence of seed banks for plant species in the Horqin Sandy Land; and (3) determine the relative importance of seed traits. This study provides a tool for studying the persistence of soil seed banks and could be useful in the restoration of vegetation and conservation of plant diversity in the Horqin Sandy Land.

MATERIALS AND METHODS

The study area was the Wulanaodu region located in the southeast of the Horqin Sandy Land, northeastern China (119° 39′ - 120° 02′ E; 42° 29′ - 43° 06′ N; 480 m a.s.l.). In the region, the mean daily temperature is –14.0 °C in the coldest month (January) and 23.0 °C in the warmest month (July). The mean annual precipitation is 350 mm and 70% of precipitation falls in June, July and August. The annual mean wind velocity is 4.4 m/s and the number of gale days (>16 m/s) is 21-80 (Zhang et al., 2016).

The seeds of plant species were collected in the Wulanaodu region and measured in the laboratory. Seed information was recorded, including name of species, family, ecological types, life forms, appendages, diaspore types, and seed mass and shape. Seed traits have been reported in a series of studies (Liu et al., 2003; Liu et al., 2004; Yan et al., 2004). Seed shape was measured by the variance in three dimensions: length, width and height. The seed shape values range from 0 to 1. A larger seed shape value indicates that the seed shape is closer to circular. The persistence of seed banks of these species has been recorded in a previous study (Zhao et al., 2011). The seed mass, seed shape and persistence of seed banks of 91 species constituted the dataset in this study and were used in the application of the random forest algorithm.

The seed mass and seed shape were input variables and the persistence of seed banks was the supervised variable. In the random forest algorithm, two-thirds of the data were selected as the bootstrap sample of a certain tree to classify the object. The remaining one-third of the data worked as test sets; thus, a cross-validation was not needed. Each tree gives a classification and the forest chooses the classification given by the most trees (Breiman, 2001).

The accuracy of the model was calculated using the OOB (out-of-bag data) error estimate. The relative importance of

the input variables, i.e. seed mass and seed shape, was analyzed using the mean decrease in accuracy and mean decrease Gini. The mean decrease in accuracy is equal to the increase in MSE (mean-squared error) and the mean decrease Gini measures the impurity of nodes in decision trees (Wu et al., 2008). To assess the partial (marginal) effects of seed mass and seed shape on the persistence of seed banks, a partial dependence plot was used. A partial dependence plot provides a graphical depiction of the marginal effect of a variable on the classification probability (Friedman, 2001). In the partial regression plot, the x-axis represents the independent variable (X_i) and the y-axis represents half of the logits for the k class (k = 1, 2 in this study). The random forest algorithm and other calculations were conducted using the "randomForest" packages in R software (version 3.2.4, R Development Core Team, 2009).

RESULTS

The seed mass of species ranged from 0.05 to 130.8 mg, and the 25th percentile, median and 75th percentile values were 0.41, 0.99 and 3.21 mg, respectively. For the seed mass data, the number of points whose value was larger than the upper whisker values was 14 (Fig. 1). The seed shape of species ranged from 0.007 to 0.207, and the 25th percentile, median and 75th percentile values were 0.05, 0.088 and 1.333, respectively. The number of species with and without persistent seed banks was 52 and 39, respectively.

For the species with a persistent seed bank, 39 were classified correctly and 13 were classified falsely, indicating that the class error rate of the existence of persistent seed banks was 25%. Meanwhile, for the species without persistent seed banks, 22 were classified correctly and 17 were classified falsely, indicating that the class error rate of the absence of persistent seed banks was 43.6% (Table 1). The OOB estimate of error rate was 32.97%.

The mean decrease in accuracy and mean decrease Gini of seed mass were 18.26 and 22.48, respectively. The mean decrease accuracy and mean decrease Gini of seed shape were 9.90 and 21.21, respectively (Fig. 2).

With increasing seed mass, the log of P (where P is the ratio of the number of votes selecting existence of a persistent seed bank to the number of votes selecting absence of a persistent seed bank) gradually decreased (Fig. 3). The log of P was above 0 when the seed mass was less than 10 mg, and was below 0 when the seed mass was more than 10 mg (Fig. 3).

The log of P decreased as the seed shape decreased in the range 0.07 to 0.17. When the seed shape was less than 0.07 and more than 0.17, the log of P increased. The log of P was equal to 0 or greater than zero when the seed shape was in the range 0.02 to 0.15 (Fig. 4).

Table 1. The confusion matrix in the persistence of seed banks. The rows indicate real classification; the columns indicate predicted classification. The existence of a persistent seed bank is represented by "1"; the absence of a persistent seed bank is represented by "0".

Tabla 1. Matriz de confusión en la persistencia de bancos de semilla. Las filas indican la clasificación real: las columnas indican la clasificación predictiva. La existencia de un banco de semillas persistente es representado por "1"; la ausencia de un banco de semillas persistente es representado por "0".

	0	1	Class error rate (%)
0	22	17	43.6
1	13	39	25

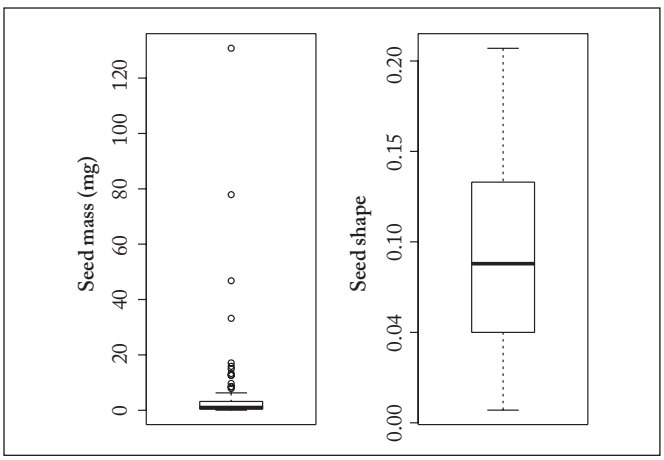

Fig. 1. Boxplots of seed mass and seed shape.
Fig. 1. Diagrama de cajas para la masa y la forma de las semillas.

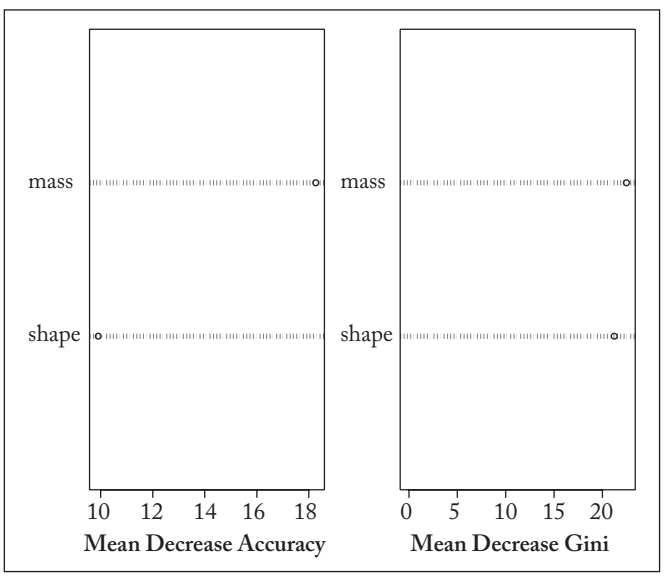

Fig. 2. Relative importance of variables with the random forest algorithm.
Fig. 2. Importancia relativa de las variables con el algoritmo del bosque al azar.

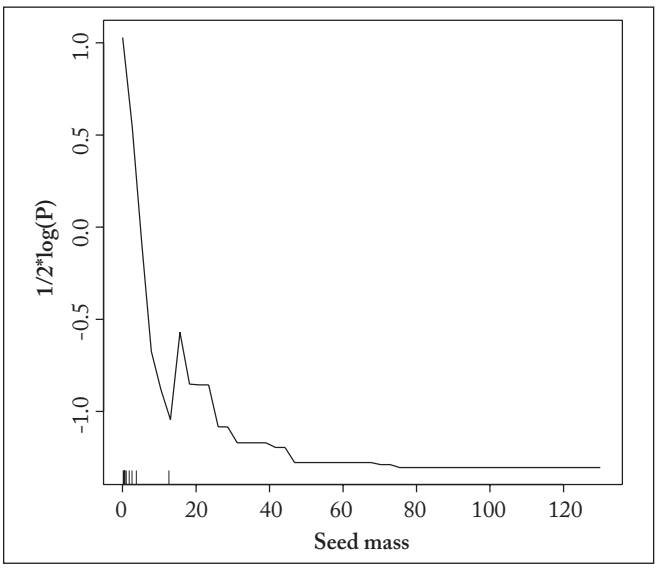

Fig. 3. Partial dependence plot of seed mass. P indicates the ratio of the number of votes selecting existence of a persistent seed bank to the number of votes selecting absence of a persistent seed bank.

Fig. 3. Figura de dependencia parcial de la masa de semillas. P indica la relación del número de votos que seleccionan por la existencia de un banco de semillas persistentes con respecto al número de votos que seleccionan por la ausencia de un banco de semillas persistentes.

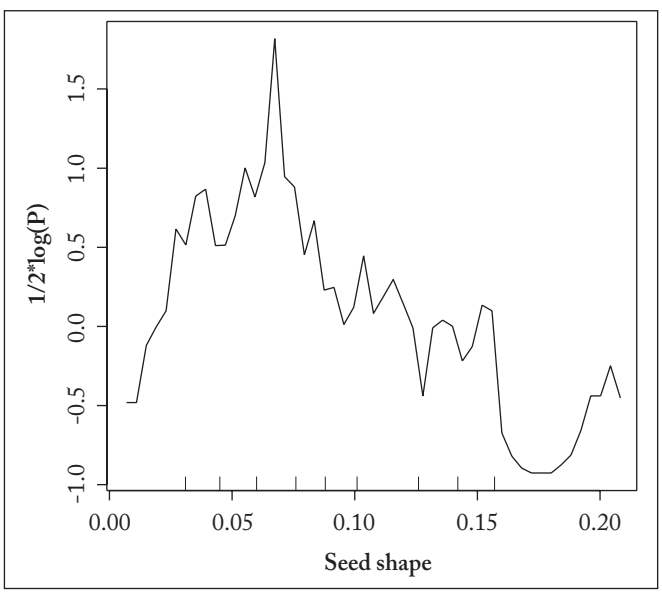

Fig. 4. Partial dependence plot of seed shape. P indicates the ratio of the number of votes selecting existence of a persistent seed bank to the number of votes selecting absence of a persistent seed bank.

Fig. 4. Figura de la dependencia parcial de la forma de la semilla. P indica la relación del número de votos que selecciona por la existencia de un banco de semillas persistentes con respecto al número de votos que selecciona por la ausencia de un banco de semillas persistentes.

DISCUSSION

In this study, the OOB estimate of error rate was 32.97%, indicating that the predictive accuracy of the random forest algorithm in predicting the persistence of seed banks was 67.03%. Seed shape and seed mass, and other traits not mentioned in this study, were used to predict the persistence of seed banks. In fact, seed persistence is not just determined by seed size and seed mass. Other factors, such as germination requirements, dormancy and resistance to pathogens, could also contribute to seed persistence in soils (Thompson et al., 1993b). Thus, using more seed traits as input variables might increase the predictive accuracy of the random forest algorithm.

The random forest algorithm has several advantages. For instance, it is not sensitive to multicollinearity, can deal with thousands of input variables and does not need to select variables, and is robust with unbalanced data (Li & Wang, 2013). These advantages mean that it could be widely used to predict the persistence of seed banks in varied environments and diverse communities.

The mean decrease in accuracy for seed mass and seed shape were 18.26 and 9.90, respectively, suggesting that seed mass is relatively more important for predicting the persistence of seed banks. This result is consistent with a previous study, in which it was claimed that seed mass is a better predictor than seed shape when predicting the persistence of seed banks in soil (Zhao et al., 2011). Compared with seed shape, seed mass is closely related to seed dormancy and variations in reproductive success, which could lower mortality risk and promote species coexistence (Rees, 1993; Pake & Venable, 1996). Furthermore, seed mass is vital for resistance to the disadvantageous habitats created by sand-blown activities in the Horqin Sandy Land (Yan et al., 2005a; Tang & Liu, 2012).

With increasing seed mass, the log of P changed from positive to negative, indicating that small-seeded species tend to have a persistent seed bank and larger-seeded species tend to germinate in the current year. This result is consistent with previous studies conducted in Argentina (Funes et al., 1999) and China (Zhao et al., 2011). Furthermore, when the seed mass was about 10 mg, the log of P was 0, suggesting that the probability of the existence of a persistent seed bank was equal to the probability of the absence of a persistent seed bank. A seed mass of 10 mg might be the dividing point for predicting the persistence of seed banks for flora in the Horqin Sandy Land.

The log of P decreased as the seed shape decreased in the range 0.07 to 0.17, suggesting that the seed shape (0.07-0.17) is negatively related to the existence of a persistent seed bank. Furthermore, the log of P increased when the seed shape was less than 0.07 and more than 0.17, indicating that seed shape <0.07 and seed shape >0.17 are positively related to the existence of persistent seed banks. This result

is slightly different from a previous study conducted in this region. In the previous study, Pearson correlation was used to explore the relationship between seed shape and the persistence of seed banks. Pearson correlation is useful for testing linear relationships and is usually poor for exploring nonlinear relationships. Thus, in the case where a threshold is found, the random forest algorithm is better than the Pearson correlation.

The effect of seed shape on the existence of persistent seed banks is relatively complex in the Horqin Sandy Land, where the position of the seed bank could be relatively easily changed by sand-blown activities. For instance, compared with elongated or flattened seeds, round ones could be more easily moved on the ground, which might mean that the distribution of elongated or flattened seeds is deeper than round seeds (Tang & Liu, 2012). Seeds in different positions on the sand dunes might differ in germination rate. Thus, species with different shaped seeds might differ in their strategies for maintaining a persistent seed bank. This study showed that the log of P was 0 when the seed shape was between 0.02 and 0.15. This indicates that the probability of the existence of persistent seed banks is equal to the probability of the absence of persistent seed banks when the seed shape is between 0.02 and 0.15. This result suggests that the role of seed shape in the persistence of seed banks is more complex than previously thought. The effect of seed shape might be related to specific physical–ecological processes, such as sand burial and wind erosion in sandy lands (Yan et al., 2005b).

In conclusion, a persistent seed bank is helpful for plant species adapting to specific habitats in the Horqin Sandy Land. To predict the persistence of seed banks, the random forest algorithm is a useful tool. Furthermore, seed mass is a better predictor than seed shape. This result confirms that small-seeded species tend to have a persistent seed bank. The role of seed shape in the persistence of seed banks is more complex than originally thought and needs to be further explored. This study expands the application of the random forest algorithm and might promote studies on persistent seed banks, especially in sandy lands.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (31870709, 41301205), Liaoning Social Science Foundation (L17BGL009) and Shenyang Science and Technology project (18013005).

REFERENCES

Ariki, S. & I. Washitani (2000). Seed dormancy/germination traits of seven *Persicaria* species and their implication in soil bank strategy. *Ecological research* 15: 33-46.

Borgy, B., X. Reboud, N. Peyrard, R. Sabbadin & S. Gaba (2015). Dynamics of weeds in the soil seed bank: A hidden Markov Model to estimate life history traits from standing plant time series. *PLOS ONE* DOI:10.1371/journal.pone.0139278.

Breiman, L. (2001). Random forests. *Machine Learning* 45: 5-32.

Crego, R.D., K.A. Didier & C.K. Nielsen (2014). Modeling meadow distribution for conservation action in arid and semi-arid Patagonia, Argentina. *Journal of Arid Environments* 102: 68-75.

Friedman, J.H. (2001). Greedy function approximation: the gradient boosting machine. *Annual of Statistics* 29: 1189-1232.

Funes, G., S. Basconcelo, S. Diaz & M. Cabido (1999). Seed size and shape are good predictors of seed persistence in soil in temperate mountain grasslands of Argentina. *Seed Science Research* 9: 341-345.

Li, H. & Y. Wang (2013). Applying various algorithms for species distribution modeling. *Integrative Zoology* 8: 124-135.

Li, J., A.D. Heap, A. Potter & J.J. Daniell (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software* 26: 1647-1659

Liu, C., M. White, G. Newell & P. Griffioen (2013). Species distribution modelling for conservation planning in Victoria, Australia. *Ecological Modelling* 249: 68-74.

Liu, Z., Q. Yan, B. Liu, J. Ma & Y. Luo (2007). Persistent soil seed bank in *Agriophyllum squarrosum* (Chenopodiaceae) in a deep sand profile: Variation along a transect of an active sand dune. *Journal of Arid Environments* 71: 236-242.

Liu, Z., Q. Yan, C.C. Baskin & J. Ma (2006). Burial of canopy-stored seeds in the annual psammophyte *Agriophyllum squarrosum* Moq. (Chenopodiaceae) and its ecological significance. *Plant and Soil* 288: 71-80.

Liu, Z., R. Li & X. Li (2004). A comparative study of seed weight of 69 plant species in Horqin sandy land, China. *Acta Phytoecologica Sinica* 28: 225-230.

Liu, Z., X. Li & R. Li (2003). A comparative study on seed shape of 70 species in Horqin sandy land. *Acta Prataculturae Sinica* 12: 55-61.

Ma, J. & Z. Liu (2008). Spatiotemporal pattern of seed bank in the annual psammophyte *Agriophyllum squarrosum* Moq. (Chenopodiaceae) on the active sand dunes of northeastern Inner Mongolia, China. *Plant and soil* 311: 97-107.

Pake, C.E. & D.L. Venable (1996). Seed banks in desert annuals: implications for persistence and coexistence in variable environments. *Ecology* 77: 1427-1435.

R Development Core Team (2009). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org.

Rees, M. (1993). Trade-offs among dispersal strategies in British plants. *Nature* 366: 150-152.

Saatkamp, A., L. Affre, T. Dutoit & P. Poschlod (2009). The seed bank longevity index revisited: limited reliability evident from a burial experiment and database analyses. *Annals of Botany* 104: 715-724.

Schutte, B.J., A.S. Davis, S.A. Peinado Jr & J. Ashigh (2014). Seed-coat thickness data clarify seed size–seed-bank persistence trade-offs in *Abutilon theophrasti* (Malvaceae). *Seed Science Research* 24: 119-131.

Tang, Y. & Z. Liu (2012). Advances, trends and challenges in seed bank research for sand dune ecosystems. *Chinese Journal of Plant Ecology* 36: 891-898.

Tang, Y., D. Jiang & X. Lv (2014). Effects of exclosure management on elm (*Ulmus pumila*) recruitment in Horqin Sandy Land, Northeastern China. *Arid Land Research and Management* 28: 109-117.

Thompson, K. (1993a) Persistence in soil. In: Hendry, G.A.F. & Grime, J.P. (eds.), pp. 199-202. Methods in comparative plant ecology: A laboratory manual. London, Chapman & Hall.

Thompson, K., S.R. Band & J.G. Hodgson (1993b). Seed size and shape predict persistence in Soil. *Functional Ecology* 7: 236-241.

Wu, X., V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. Mclachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D.J. Hand & D. Steinberg (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems* 14: 1-37.

Yan, Q., Z. Liu & R. Li (2005a). A review on persistent soil seed bank study. *Chinese Journal of Ecology* 24: 948-952.

Yan, Q., Z. Liu, J. Zhu, Y. Luo, H. Wang & D. Jiang (2005b). Structure, pattern and mechanisms of formation of seed banks in sand dune systems in northeastern Inner Mongolia, China. *Plant and soil* 277: 175-184.

Yan, Q., Z. Liu, Y. Luo & H. Wang (2004). A comparative study on diaspore weight and shape of 78 species in the Horqin steppe. *Acta Ecologica Sinica* 24: 2422-2429.

Yu, S., H. Chen & N. Lang (2007). The classification systems of soil seed banks and seed persistence in soil. *Acta Ecologica Sinica* 27: 2099-2108.

Zhang, M., J. Wu &Y. Tang (2016). The effects of grazing on the spatial pattern of elm (*Ulmus pumila* L.) in the sparse woodland steppe of Horqin Sandy Land in northeastern China. *Solid Earth* 7: 631-637.

Zhao, L., G. Wu & J. Cheng (2011). Seed mass and shape are related to persistence in a sandy soil in northern China. *Seed Science Research* 21: 47-53.