Asian Journal of Water, Environment and Pollution, Vol. 19, No. 3 (2022), pp. 89-95.
DOI 10.3233/AJW220044

Creating a Soil Map Using Digital Soil Mapping on the Example of the Diurtiulinsky Municipal District

Elina Shafeeva*, Ilnur Miftakhov, Marat Ishbulatov and Oleg Lykasov

Department of Real Estate Cadastre and Geodesy, Federal State Budgetary Educational Establishment of Higher Education "Bashkir State Agrarian University", Ufa, Russian Federation
✉ shafeeva_el@rambler.ru

Received August 2, 2021; revised and accepted April 15, 2022

Abstract: The study describes the use of machine learning methods, geostatistics, etc. in establishing soil properties depending on various classes of soil. The most commonly used data are information from spectral reflectance bands of satellite images and terrain models. Besides, there is also great potential for creating new data tiers. The study relies on a method known as SCORPAN-SSPFe, which assumes spatial error autocorrelation as a standalone function. This method is actively used in places where there is not enough information about soil data. Besides, four types of interpolation were compared using the SCORPAN method: multiple linear regression, cubistic model, cubistic model with kriging and random forest model, which use extensive but common values of soil properties associated with soil classes. The research result is obtained by applying the method to conduct large-scale soil surveys, which determines the purpose and relevance of our study.

Key words: Soil mapping, GIS technologies, geostatistical modelling, geoinformation system, methods of digital soil mapping.

Introduction

Soil is involved for efficient agricultural production and food security, therefore, contributing to the conservation of biodiversity and pollution buffering. Soil parameters as input data are used for numerous ecological and socio-economic models to assess and predict future environmental changes. It is necessary to conduct a soil survey to obtain information on resources in order to satisfy not only science but also consumers. An alternative is a modern research, which, unlike classical studies, requires less cost and time, and is also less financially limited when mapping soils on a large scale. Webster and Oliver (2016), in their study "Modeling spatial soil variability as random functions", describe the geostatistical approach used to characterise the heterogeneity of soil cover and the diversity of soil properties. Researchers point out the shortcomings of classical soil mapping, which has stimulated soil scientists to develop modern methods for obtaining soil data. Steinbuch et al. (2019), in their study, "Model-Based Geostatistics from a Bayesian Perspective: Investigating Area-to-Point Kriging with Small Data Sets", considered a geostatistical method for creating high-resolution raster maps using data from a variable of interest with a much lower resolution, which is called Area-to-point kriging (ATPK). There are two main approaches to predicting the state of the soil in hard-to-reach places. One of them consists in dividing the soil cover into spatial units, within which the soils correspond to the features of the soil classification class.

The second approach considers soils as a set of continuous variables describing the change of these variables depending on the landscape. The

*Corresponding Author

second approach is quantitative for interpolation established between soil survey points. All soil research programmes conducted around the world during the 20th century were based on the first approach. Arnold (2006), in his study the "Soil survey and classification of soils", described the methods and procedures necessary for making a soil map using this approach (Arnold, 2006). In the first stage of the study, the soil scientist observes the soil's physical properties to correlate the soil characteristics with the landscape features. Aerial photographs and topographic maps are often used as auxiliary information. This was helpful to describe the type of soil. The second stage is creating a conceptual model, based mainly on the landscaping knowledge of a soil scientist, which depends on his skills and previous experience. The third stage involves transferring the conceptual model to the cartographic model, usually based on an orthophoto with a plotted relief. A conceptual model is created by delineating a territory that has the same soil-landscape ratios. All these stages are described by Hewitt et al. (2015) in their research.

By the end of the 20th century, there was enough knowledge in using mathematical and statistical research methods in ecology and soil science. In 2013, Webster and Oliver presented their book "Geostatic Analysis for Environmental scientists" (Webster and Oliver, 2013). Currently, some statistics packages have the necessary tools to fulfil these purposes. Soil maps compiled using traditional approaches have some.

Minasny et al. (2013) generalised various approaches to quantitative soil mapping and developed a structure called SCORPAN (**S**: Soil, **C**: Climate, **O**: Organisms, **R**: Terrain, or topography, **P**: Source material, including lithology, **A**: Age, **N**: Space). This structure is not used for a simple explanation but rather for numerical description of soil-landscape areas to make spatial forecasts and create more efficient soil maps within the new paradigm of digital mapping.

The SCORPAN structure develops the ideas of Minasny et al. (2013). It is formally known as SCORPAN-SSPFe and includes soil spatial prediction function (SSPF) and auto-correlated errors (Minasny et al., 2013).

Some researchers believe that soil properties, which are significant factors, can be predicted based on their spatial arrangement. An example of such a prediction is kriging. Prediction is also possible based on the distance to significant landscape features: hilltops, rivers, roads, pollution sources, etc. (Malone, 2018).

The logic of digital soil mapping (DSM) is based on the application of constraint equations, which Minasny et al. (2013) described for this mapping method.

The formation of the theory of statistics contributes to the constant development of the applied mathematical models and the creation of new ones. The study considers them in a more detailed manner. Continuous variables are more often modeled using well-known tools such as regression trees or multiple linear regression. For ordinal or categorical data, classification trees or logistic regression are predominantly used.

Among the simplest modeling methods, it is worth mentioning the linear models using generalised or standard fit forms in the framework of the least squares method. There are also more complex models that generalise linear models or use generalised additive types of models. According to Kempen et al. (2019), models based on logistic regression are only a form of generalisation of linear models that successfully model categorical type variables.

Regression can be used to process any residual spatial autocorrelation that is probably the result of fitting the SCORPAN model. DSM often uses regression-kriging approach, the generalised form of which is a combination of computer-added learning methods and classical geostatistical residual modeling (Coca-Castro et al., 2021).

According to Kempen et al. (2019), the well-known variant for the type of soil categorical prediction for the creation of DSM types (multinomial logistic regression, classification trees, etc.) is non-spatial models since, unlike continuous variables, the spatial properties of the target variable are not considered (Kempen et al., 2019). The same researchers were engaged in a linear geostatistical model, with the help of which they had previously tried to solve the described problem. On the one hand, the method studied by the scientists was geostatistically attractive. On the other hand, it was tedious and did not significantly improve accuracy compared to the non-spatial polynomial logistic regression model.

Chai (2015) presented and described the method of reference vectors as a method of digital mapping of soils and soil properties.

The DSM study includes many proven potential models. One example is the R statistical computation language and the Carpet software package for its processing as applied in Kuhn et al. (2016) study. As of March 2017, 448 modelling algorithms are available for continuous and categorical target variables (Kuhn et al. 2016). Nevertheless, many of them will not be suitable

for the DSM, and in the future, the Carpet package will continue to add new algorithms.

Modern digital soil mapping has not yet been thoroughly studied and requires development from the point of view of the scientific society (Zhang, 2017).

Study Area

The study object is the physical and geographical conditions of soil formation in the Republic of Bashkortostan in the example of the Diurtiulinsky municipal district (Southern forest-steppe part).

The research area is to the northwest of the Republic of Bashkortostan along the lower course of the Belaia River. The region is characterised by a warm and slightly arid climate and is famous for relict pine and spruce forests (Iaparov, 2005).

Methods

The soils are mainly typical and leached chernozems, but podzol grey forest and floodplain soils also exist. The shape of the SSPF is usually defined at the very beginning of a project, based on DSM. The SCORPAN model is based on formula (1):

$$Sc = f(s,c,o,r,p,a,n) \text{ and } Sp = f(s,c,o,r,p,a,n)$$

In the formula Sc presents soil classes and Sp are soil properties or attributes.

Information representing other SCORPAN factors is described in Table 1.

Table 1: Possible sources of information for representing the SCORPAN function

SCORPAN functions	*Possible representatives*
S	Outdated soil maps, point observations, expert knowledge
C	Temperature and precipitation records
O	Vegetation maps, species abundance maps, yield maps, land use maps
R	Digital terrain model, terrain attributes
P	Outdated geological maps, gamma-ray radiometric information
A	Weathering indices, geological maps
N	Latitude and longitude or direction to the east and north, distance from landscape features, distance from roads, distance from point sources of pollution

Figure 1 shows the scheme of DSM carried out using the method described in the book of Boettinger (2010)

The main methods of the study are:

1. Multiple linear regression;
2. Cubist models;
3. Cubistic models with kriging of model residuals;
4. Random Forest models

While collecting material for the study area, a soil map of the first detailed soil survey was obtained.

The obtained soil maps of the farms served as source material for a map of the area as a whole. The original soil map is shown in Figure 2.

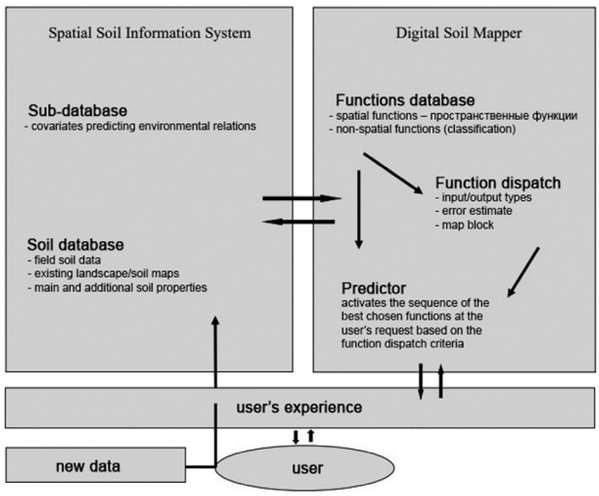

Figure 1: Diagram of DSM creation.

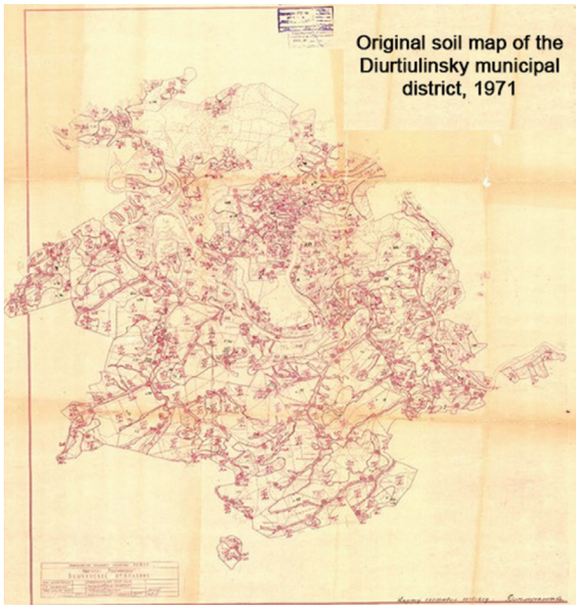

Figure 2: Original soil map of the Diurtiulinsky municipal district, 1971.

According to the technical reports and the soil map of 1971, the territory of the investigation area is represented by 17 soil types and 75 varieties. During the study, a digital soil map of the 2019 survey was used. The map was created by specialists of JSC "VolgoNIIgiprozem" using traditional methods.

As for the models considered, the multiple linear regression is concerned with modeling the relationships between the response and explanation variables, which is accomplished by fitting a function. Filtering out variables that may not have a significant impact on the performance of the model is most commonly done with stepwise regression.

Consideration of the non-linear form of relationships between the studied data requires the use of special analysis algorithms, for example, the cubist model. It is typologically close to the regression tree according to the separation algorithm used for data. This model decomposes the data using recursion, creating subsets that are much more homogeneous with respect to the target covariates than the entire dataset under investigation. Subsets are characterized by a set of rules that constitute their own hierarchy.

Cubistic models with residual kriging are an improvement on using only cubistic modelling. In this case, after training the cubistic model, the model residuals (the difference between a set of observations and the forecast of the model) are saved to evaluate if they demonstrate their spatial pattern (autocorrelation). The evaluation is carried out by fitting the variogram to the residual data. Kriging begins immediately after the completion of the model candidate selection process. Variogram fitting can be done in the local area for a small dataset or globally for the entire dataset, which usually depends on the volume of data available. Ultimately, the output of the regression model and the krigged residuals merged to produce the conclusive forecast. Finally, random forests are a reinforced approach to the decision tree. Boosting is done using the ensemble learning method based on building decision set trees aggregated to give the only forecast for each dataset observation. A reasonable idea of a good fitting of the training model can be formed with the help of so-called "ready fetches", which are data that are internally held during the construction of the tree. More detailed information about the popular model can be found in the studies of Minasny et al. (2013), which are an example of its application when making a DSM.

Statistical data processing and modelling were carried out using the R software environment, programs MS EXCEL-2007 and STATISTICA 10.0.

Results

The area of the study place is 117,229.17 hectares. According to the complexity of large-scale soil surveys, the area belongs to category 2. Thus, drawing up a map of soil at the scale of 1: 50,000, one section per 130 hectares is required. The ratio between soil sections, semi-pits and by-pits should be 1:4:5 based on topographic maps.

Consider an example that demonstrates a typical workflow for DSM. For illustrating this workflow, several different spatial prediction functions should be applied. These functions demonstrate just some models suitable for the DSM. In these examples, the laboratory parameters of the soil sections serve as the target variable. The soil data set includes 210 observations collected in the Diurtiulinsky municipal district. The mapping covers a small area for simplicity (approximately 117,229.17 ha) of the district where the samples were taken. As functions for predicting the spatial position in this example, we used (1) multiple linear regression, (2) cubistic models, (3) cubistic models with kriging model residuals, and (4) random forest models. Figure 3 shows the distribution of sampling points. Several spatial datasets of spatial positions were collected for the region under study. All of them were recorded, considering their original resolution and grid cell distances of 25 m. All of them are used as elements of forecast and environmental covariates.

Here, the data mainly demonstrates the factors **of r** and **o** for the SCORPAN structure. The representation of **r** factor variables was extracted from a digital terrain model, while the variables showing **o** values extracted from remote sensing images obtained using the Landsat 7 ETM+satellite platform.

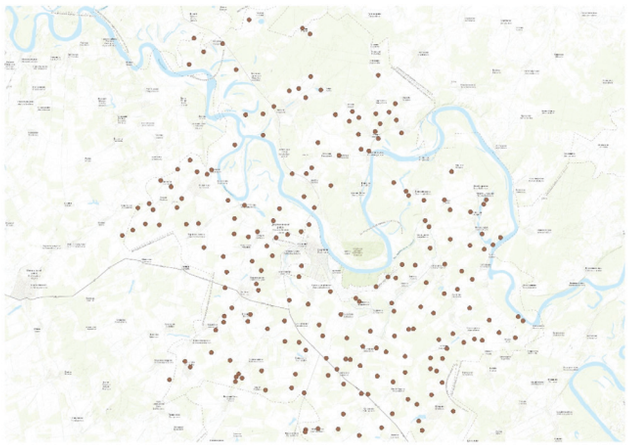

Figure 3: Distribution of survey points.

Quantifying the functions of spatial forecast effectiveness is a distinctive feature of a digital soil map. According to Kaladhar et al. (2013), common statistics model suggests the mean square error (RMS error), the average error (model bias), the determination coefficient (CD), and the concordance correlation coefficient. These statistics should be calculated using an independent test dataset to verify their objectivity, evaluate the effectiveness of the model generalisation, and make recommendations for choosing the optimal model.

The experimental variogram model was used in the study. Segments connect the averaged points, and the resulting polyline is called an experimental variogram. The experimental variogram gives a considerable amount of information about the behaviour of a spatial random variable, namely, the variables of the SCORPAN method. The parameters of the variogram and the effect of the method can be estimated from the empirical variogram. An unbiased estimate of the variogram function is half of the standard deviation between the values of data pairs.

The relationship between the target variable and the spatial data is studied at the model fitting stage. Once implemented, the parameters allow the entire display extent to be covered by target variable prediction. Figure 4 demonstrates the digital soil maps obtained by applying each model over the entire display extent. From Figure 4, one can observe the subtle maps differences, but they demonstrate identical spatial pattern. Please note that this may not correspond to observations on other DSM models. Table 2 summarises the RMS error and CCC statistics for each model with regard to training and validation data sets. This is usually the case when the quality of fitness seems to be better for training compared to validation.

Table 2: Compliance statistics for spatial models based on training and validation datasets. Concordance correlation coefficient (CCC); root-mean-square error (RMS error)

Parameters	*Training*		*Definition*	
	CCC	*RMS error*	*CCC*	*RMS error*
Multiple linear regression	0.40	1.17	0.40	1.19
Cubist model	0.47	1.13	0.41	1.19
The cubistic model with residual kriging	0.85	0.66	0.60	1.05
Random forest model	0.89	0.53	0.32	1.22

Table 2 shows the minor differences between multiple linear regression and the cubist model, which indicates that both models are not very suitable or are applicable in general conditions. Cubistic models with residual kriging provide the forecast of the best validity. The random forest model provides the worst predictions, but it is the best in training data. The example clearly shows the importance of the model's compliance evaluation using both validation and training data. The random forest model seems very accurate, but in fact, it is quite susceptible to overfitting.

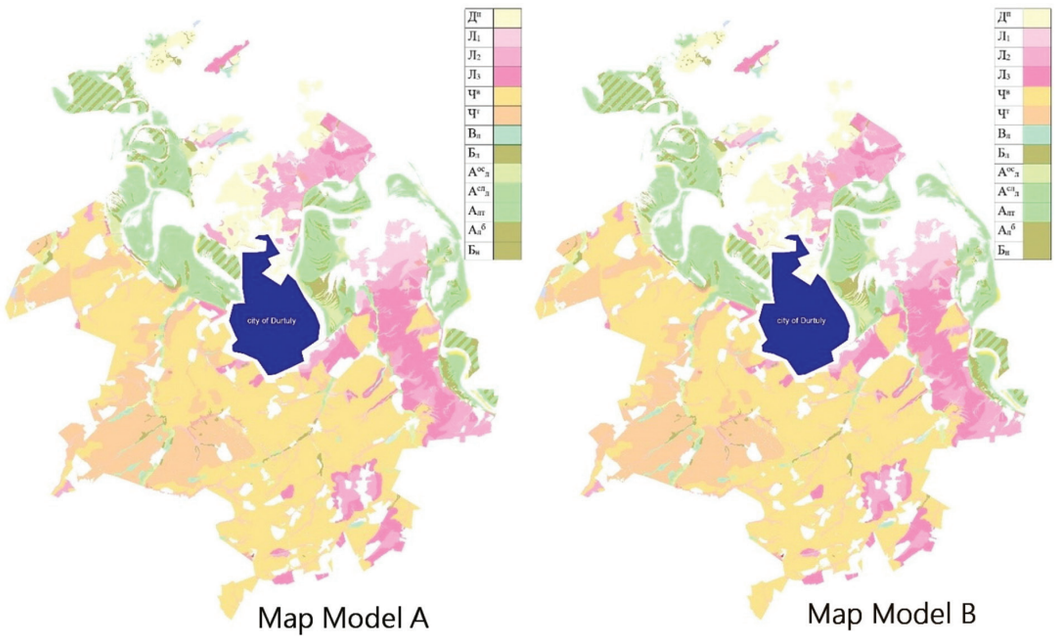

Figure 4: Mapping as a result of multiple linear regression (A), cubistic modeling (B).

Discussion

Most of the methods to create a DSM are based on the use of observable point soil data combined with environmental covariates and the model function for spatial prediction. However, it is not limited to such natural datasets. A set of previous components of disaggregated soil map were made when map units were disaggregated into classes or series, the spatial structure is defined by the function of the spatial forecast. Once the researcher has identified an area of interest and collected a set of ecological covariates for that area, depending on the available data, approaches that can be realised to create digital soil maps are proposed. The traditional approach is kriging scorpan. It is used if it is possible to obtain data of an exclusive point nature. It is also applicable in the case of the availability of not only point, but also more accurate cartographic data. Scorpan kriging is complemented by disaggregation of soil maps based on a combinatorial approach using an averaged model. Odgers et al. (2015) conducted such studies in Australia. The scale and subsequent changes in the soil cover directly affect the image quality of soils. Thus, maps of a smaller scale, such as 1: 100,000, will be more detailed and clearer to understand than maps of a larger scale, such as 1: 500,000. Taking soil properties data from inherited images of soil maps is based on soil cartographic units' central and distributional concepts. For example, the modal data on the soil profile of soil classes can be used to build the properties cartograms of soil quickly. An early example of this is the multi-layer set of soil characteristics for external soil data (CONUS-SOIL) used in various climates, hydrology research, and models of the landscape.

Conclusion

This research uses exhaustive covariates and modeling methods to build the most reliable and accurate model for estimating the spatial spreading of soil for the MR Dyurtyulinsky district. The findings demonstrate that qualitative environmental variables can develop the accuracy of soil distribution estimates; in particular, each category of qualitative variables showed significant differences with soil values. The obtained results also demonstrated the importance of the choice of variables in modeling. Topography and organisms are the most critical ecological soil-forming factors in soil forecasting. After selecting the variables, the most important environmental variables explaining the observed variability in soil thickness were slope, height, and land use. The Random Forest Model achieved the best performance of the four individual models, followed by the rest. However, compared to these different methods, the two stacking models showed better performance. The best model (the Random Forest Model) showed the highest performance with the lowest average square error (0.53). Judging by the maps of the spatial soil spreading, the soils were mainly concentrated in valleys and low altitudes, and shallow soils were mainly found on steep slopes and high altitudes. The study has the potential to be applied in the practice of creating large-scale soil maps. When using model algorithms for digital soil mapping, selecting the appropriate variables and models is recommended.

References

Arnold, R.W. (2006). Soil survey and soil classification. Environmental soil-landscape modeling: Geographic information technologies and pedometrics. Taylor & Francis Group, Boca Raton, FL, 37-60 pp.

Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E. and S. Kienast-Brown (Eds.). (2010). Digital soil mapping: Bridging research, environmental application, and operation. Springer Science & Business Media.

Chai, H., Rao, S., Wang, R., Liu, J., Huang, Q. and X. Mou (2015). The effect of the geomorphologic type as surrogate to the time factor on digital soil mapping. *Open Journal of Soil Science*, **5(06):** 123.

Coca-Castro, A., Gutierrez-Díaz, J.S., Camacho, V., López, A.F., Escudero, P., Serrato, P.K. and J. González (2021). Optimized data-driven pipeline for digital mapping of quantitative and categorical properties of soils in Colombia. *Revista Brasileira de Ciência do Solo*, **45:** e0210084.

Hewitt, A., Dominati, E., Webb, T. and T. Cuthill (2015). Soil natural capital quantification by the stock adequacy method. *Geoderma*, **241:** 107-114.

Iaparov, I.M. (2005). Atlas of the Republic of Bashkortostan. Kitap, Ufa.

Kaladhar, D.S.V.G.K., Pottumuthu, B.K., Rao, P.V.N., Vadlamudi, V., Chaitanya, A.K. and R.H. Reddy (2013). The elements of statistical learning in colon cancer datasets: Data mining, inference and prediction. *Algorithms Research*, **2(1):** 8-17.

Kempen, B., Yigini, Y., Viatkin, K., de Jesus, J. M., de Sousa, L.M., van den Bosch, R. and R. Vargas (2019, January). The Global Soil Information System (GloSIS)–concept and design. *In:* Geophysical Research Abstracts (Vol. 21).

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y. and

C. Candan (2016). Caret: Classification and regression training. R Core Team. Available from: http://www.R-project.org/.

Malone, B.P., Odgers, N.P., Stockmann, U., Minasny, B. and A. McBratney (2018). Digital mapping of soil classes and continuous soil properties. *In:* Pedometrics (pp. 373-413). Springer, Cham.

Minasny, B., McBratney, A.B., Malone, B.P. and I. Wheeler (2013). Digital mapping of soil carbon. *Advances in Agronomy*, **118:** 1-47.

Steinbuch, L., Orton, T. and D. Brus (2019). Model-based geostatistics from a Bayesian perspective: Investigating area-to-point kriging with small data sets. *Mathematical Geosciences*, **52:** 1-27.

Odgers, N.P., McBratney, A.B. and B. Minasny (2015). Digital soil property mapping and uncertainty estimation using soil class probability rasters. *Geoderma*, **237:** 190-198.

Webster, R. and M.A. Oliver (2013). Geostatistics for Environmental Scientists, second edition. John Wiley & Sons, Chichester.

Webster, R. and M.A. Oliver (2016). Modeling spatial variation of soil as random functions. *In:* Environmental soil–landscape modeling: Geographic information technologies and pedometrics. CRC Press, Boca Raton, FL, 241-289 pp.

Zhang, G.L., Feng, L.I.U. and X.D. Song (2017). Recent progress and future prospect of digital soil mapping: A review. *Journal of Integrative Agriculture*, **16(12):** 2871-2885.