Geotechnical and Geophysical Site Characterisation 5 – Lehane, Acosta-Martínez & Kelly (Eds)
 ISBN 978-0-9946261-2-7

Automatic methodology to predict grain size class from dynamic penetration test using neural networks

C. Sastre, M. Benz & R. Gourvès
Sol Solution Géotechnique Réseaux, Riom, France

P.Breul & C.Bacconnet
Université Blaise Pascal, Laboratoire de Génie Civil Polytech, Clermont-Ferrand, France

ABSTRACT: The Panda 2®, developed by Roland Gourvès in 1991, is a lightweight dynamic cone penetrometer. It provides the dynamic cone resistance (q_d) and depth in real time with a high sampling frequency. Nevertheless it cannot take soil samples so the penetration test is called 'blind'. The aim of this paper is to propose an automatic methodology to predict the soil grading from the cone resistance using artificial neural networks. We have built a database based on the Panda® laboratory tests on soil samples and on in situ tests conducted next to boreholes during various geotechnical studies performed in France. Then the neural network was used to classify the cone resistance logs according to grain size distributions of the tested soils by means of feature extraction using different signal analysis. The results show that we are able to separate 4 soil classes with 98% accuracy.

1 THE PANDA 2®, VARIABLE ENERGY DYNAMIC CONE PENETROMETER (DCP)

The dynamic penetrometer Panda 2® has been designed to geotechnical investigation at shallow depth up to about 5 meters (Benz 2009). It is a light-weight dynamic, highly portable cone penetrometer, which uses variable energy manually delivered by the blow of a normalized hammer. After each blow, the dynamic cone resistance q_d is calculated at the current depth using the Dutch formula. One of the major interest is the high acquisition resolution. Therefore the plot of the cone resistance values against the depth, the Panda penetrogram, is a rich amount of information on the stratigraphy of site and soil properties (Shahour & Gourvès 2005) with a large number of data.

Despite all the benefits provided by the Panda 2® soil samples cannot be taken during the test thus there is no information about the nature of soil. However we notice empirically that the form of the cone resistance curve might differ between different types of soil. An example shows (Fig.1), with 3 Panda® tests conducted on laboratory calibration chamber of 80 cm height. The tested soils have a different nature and granulometry but a similar density and moisture content. We can easily note the morphological differences between the 3 resistance logs. The underlying idea for the methodology described in this paper relies on this observation: the expert knowledge and perspective of an experienced engineer can detect indices in the signal form to estimate the nature of the tested soil. In this study, we have developed a model based on artificial neural network to accomplish this empiricism.

Table 1. Soil parameters for the samples (Fig.1)

Nature		DGA Silt	Laschamps Clay	Sayat Sand
GTR		A1	A2	B2
γ_h	kN/m³	17.8	17.6	17.3
γ_d	kN/m³	16.2	15.6	16.0
W	kN/m³	10.0	12.9	7.9
W_{OPN}	%	18.4	18.1	11.0
Moisture state		Dry	Dry	Dry

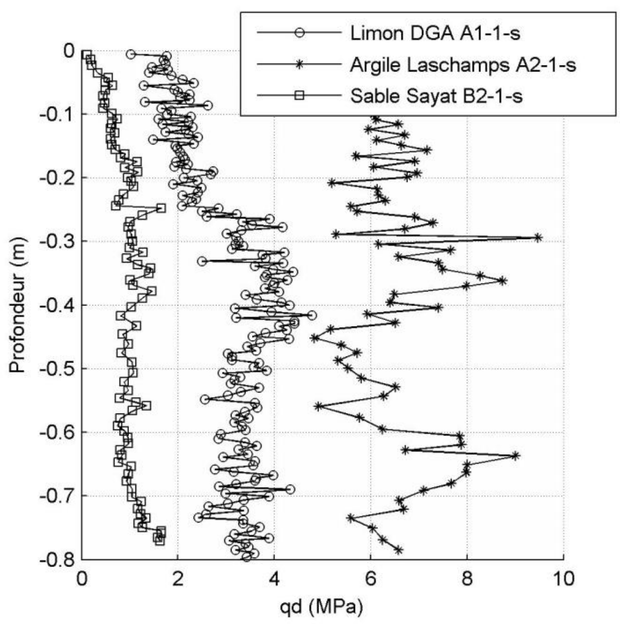

Figure 1. An example of a resistance log for different types of soils with similar state parameters.

2 ARTIFICIAL NEURAL NETWORKS

2.1 *Introduction*

ANN are mathematical models inspired in the human nervous system and has been effectively applied in many engineering applications. ANN are artificial intelligence (AI) tools to analyse the raw data and extract useful knowledge (Fayad et al. 1996) and to support selection problems. Furthermore ANN is known to be an alternative method for modelling complex problems in different fields of engineering (Shahin et al. 2001, Waszcsyszyn 2011).

2.2 *Proposed ANNs models*

Among several types of ANNs models used for data-analytic applications, multilayer perceptron MLP (Rumehart et al. 1986) and probabilistic neural network PNN (Spetch 1990) have been chosen. These models are feedforwards networks consist of multiple layers of interconnected neurons. We can distinguish 3 types of layers, input layer, hidden layer and output layer (Fig. 2). The neurons of the input layers are just used to stock the inputs values. The neurons of hidden and output layer are calculation cells. They compute a weighted sum of their inputs and then the activation function is applied to the sum in order to generate their output value. They are called feedforward networks because the input signal always moves one direction only, from input to output, and it never goes backwards.

The current training algorithm for this networks is the back-propagation method (Rumerhart et al. 1986). This algorithm involves 2 phases. Firstly the propagation phase where the training pattern's input are propagated from the input to the output layer. The second phase is the backward phase where the output error at a defined layer backwards through the connections with the previous layer in order to update the weights to find the minimum of the error function. That is known as the learning process of the ANN. The goal of the learning process is to find the optimal set of weights which would produce the right output for any input ideal case.

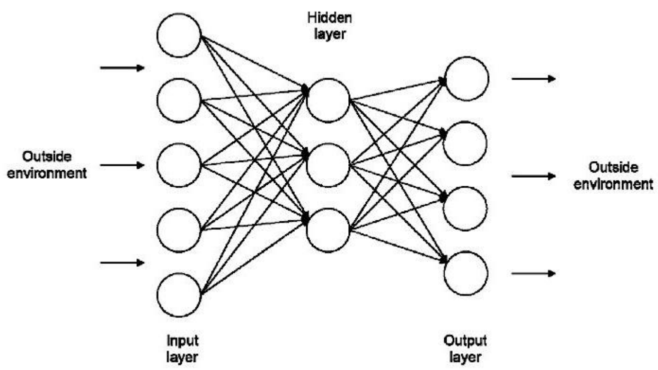

Figure 2. An example of a simple feedforward neural network.

Consequently the algorithm requires the target value for each input value to calculate the loss function gradient. It is for this reason that it is considered to be a supervised learning algorithm fundamentally. The term supervised refer to the labeled training data. In our context it may be possible to apply a supervised learning if we develop a cone resistance log data base and each sample has an appropriated soil class as output. In this way we could think about a model with an associative memory which would be able to predict the soil class from the cone resistance log based on a knowledge obtained from a labeled database.

MLP modelling has been used most often in geotechnical literature (Shahin et al. 2008). Furthermore MLP are universal approximators in the sense that they can compute an approximation that is good as we want even with only one hidden layer. On the other hand, PNN are frequently used in classification problems. They have a similar structure to MLP. The main difference is the change of the sigmoid or hyperbolic activation function often used in MLP topology by a statistically derived one, a radial base function, normally the Gaussian function. Unlike MLP, PNN approaches optimal Bayes classification and the outputs could also be used to estimate a posteriori probability that an output belongs to a defined category. On the other hand, one of the main difficulty of MLP paradigm is the fact to estimate the number of hidden layers and the number of neurons of each layer. However the PNN has always 2 hidden layers the pattern layer and the summation layer respectively. The first layer always contains one neuron for each case in the training data set and the second layer contains a number of neurons equal to number of defined targets. I

2.3 *ANNs in geotechnical problems*

In the geotechnical context, ANNs are considered as powerful modelling tools to deal with the uncertainty and extreme variability of most of the problems. In addition, ANNs have also demonstrated a major performance when compared with traditional statistical models. Therefore, since the early 1990s, ANNs have been applied successfully to several problems in geotechnical engineering (Shahin et al. 2009). Interested reader can refer to (Shahin et al. 2001, 2008, 2009) where the application of ANNs in geotechnical problems are examined.

(Sulewska 2011) pointed out the following six selected problems: 1) prediction of the Overconsolidation Ratio, 2) estimation of potential soil liquefaction, 3) prediction of foundation settlement, 4) evaluation of piles bearing capacity, 5) prediction of cohesive parameter for cohesive soils, 6) compaction built of cohesionless soils. The ANNs applied are all MLP with only one hidden layer. The number of hidden neurons varies from 1 to 8. The excellent results confirm the interest of the application of ANNs in the geotechnical field, where the choice of ANNs models to regression problems is constantly increasing.

Despite all the benefits presented by ANNs, one of the major criticisms is that they are black boxes models, since no satisfactory explanation of their behavior could be achieved so far. The knowledge extraction is one of principal research topics in this field. In addition, ANN models often need a large database to perform an effective learning of the patterns. The number of samples dependent upon the problem to solve and cannot be estimated a priori. This fact could be a constraint on the application for certain geotechnical problems.

3 PROPOSED METHODOLOGY

The aim of this study is to propose an automatic methodology able to classify a sol according to its granulometry from a cone resistance signal provided by Panda2® test. This project is divided into 4 phases: Panda test database creation; input variable selection; output model; performance model validation.

3.1 *Panda test database*

The first step in a machine learning problem is the database acquisition. The database creation is necessary to allow the ANNs models observe the environment and learn to make reasonable decisions about the categories of the patterns.

A database of penetrograms provided by the Panda2® test has been created. It contains 218 penetrograms obtained from sufficiently homogeneous soils with no grain size over 50mm. Their nature and geotechnical properties have been characterized by means of laboratory test. Samples have been provided by laboratory and in situ test. The soil classification available for the tested soils based on GTR guide (SETRA & LCP 1992).

The laboratory Panda test were realized in a 37 cm diameter and 80 cm height calibration chamber, by using static loading under oedometric conditions (Chaigneau 2000). These tests were performed in more than 20 different soil types. Panda tests were carried out for each sample constituted. The part of the penetration curve where the cone resistance exhibit constant values was extracted and submitted to the posterior signal analysis. A total of 149 laboratory test have been recollected with a dry and medium water moisture.

On the other hand, there are 69 Panda2® in situ tests performed during several site characterization studies by the French company Sol Solution. In situ test are located in France, specifically in Auvergne department.

3.2 *Feature extraction and variable selection*

The second step involves the definition of entries of ANN model. The choice of input variables is fundamental to assure the model performance. This stage is called feature extraction and mainly consists of a short-term processing technique that is applied on the observed data in order to generate a feature sequence, the pattern.

We have carried out a feature extraction based on four analysis signal applied to cone resistance curve Inspired by analysis processing for speech recognition or bioelectrical signal analysis (Shannon 1948a, b, Kannatey-Asibu 1982, Hayes 1996, Betancourt 2004, Romo et al. 2007). We have applied 4 signal analysis: statistical, nonlinear, and morphological and spectral and a pattern vector of 26 parameters for each penetrogram have been obtained.

Variable selection is intended to select the best subset of predictors, removing redundant predictors and defying the curse of dimensionality to improve classification performance. In this work we have performed a local sensitivity analysis using one -at-a-time approach (OAT). One benefit of OAT is one of the simplest approach and it can be applied to all numerical models so we have chosen this method for practical reasons. Rabitz H. (1989) and Saltelli (2006) offer an interesting review about the use of sensitivity analysis strategies for model-based inference in different articles.

OAT techniques analyze the effect of one parameter on the cost function at a time, keeping the other parameters fixed. Note that they explore only a small fraction of the design space but is enough to allow a quick detection of model inputs what don't have any significant influence in the output value. In this instance 3 perturbation values have been considerate, calculated as a percentage of the standard deviation for each input variable (Fig.3). Another perturbation equal to input variable variance have been also used. After each modification the variation of error is calculated in order to measure the individual impact of each entry parameter.

Table 2. Parametrization of cone resistance log.

Number of parameter		
1. q_d mean	7. q_d interquartile range	13. q_d slope changes
2. q_d median	8. q_d skew	14. q_d waveform
3. q_d standar deviation	9. q_d kurtosis	15. Linear coefficient of linear trend
4. q_d coefficient of variation	10. q_d Shannon entropy	16. Independent coefficient of linear trend
5. q_d variance	11. q_d logarithm entropy range	
6. q_d range	12. q_d skew	17. Maximum spectral power

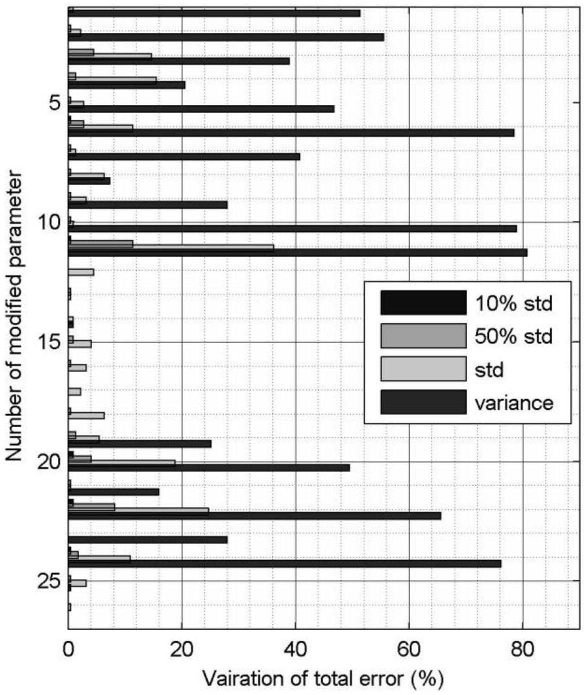

Figure 3. Results of the sensitivity analysis for the 26 original characteristics.

Analyzing the results of OAT approach used, we have realized that there are 9 inputs variables relative to cone penetration analysis which modification have no influence on the ANN accuracy classification. They are rejected and the number of input variables have been reduced to 17. In other terms, for every cone resistance log of the database, a final feature vector of 17 parameters (Tab.2) is thus constructed to be the entry for the ANN model.

3.3 *Target classes*

The next stage is to decide what outputs are the neural network expected to learn. The aim is to classify the nature soils in terms of granulometry. As we have explained in the previous section 3.1, the database soils collected are classified in GTR classification. In the following, the 4 output classes proposed (Tab.3) are based on this soil classification. We have used a binary coding, namely "dummy". Each dummy variable is given the value zero except for the one corresponding to the correct category, which is given the value one.

Table 3. Target classes based on tested soils.

Target	GTR				Nature	Codification			
Class 1	A1	A2	A3	A4	Fine soils	1	0	0	0
Class 2	B5		B6		Fine sands	0	1	0	0
Class 3	D1	B1	B2		Sands, gravel	0	0	1	0
Class 4	D2	B3	B4		Gravel	0	0	0	1

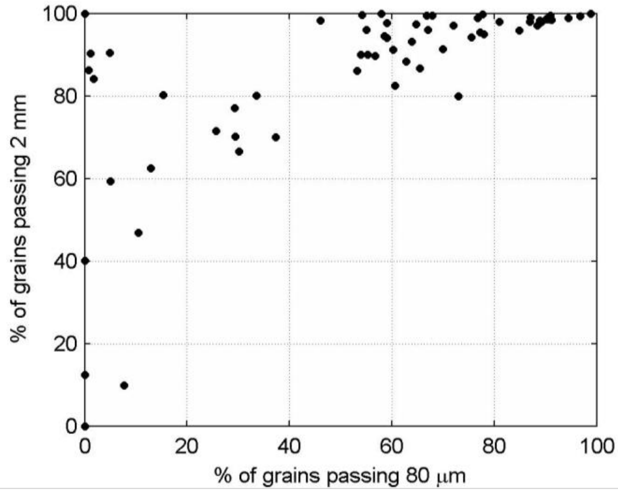

Figure 4. Percentage passing through the sieve opening 2 mm and 80 μm for the database soils.

The 2 opening size of sieves used to classify the soils in GTR are 2 mm and 80 μm. The percentage of grains passing through these openings is plotted (Fig.4) for the database soils.

3.4 *Design phase for MLP classifier*

The precise network topology required for an MLP to solve a particular problem usually cannot be determined (Leverington 2009), although research efforts continue in this regard. In contrast, PNN networks don't have this issue because their architecture is fixed by the size of the problem as we have explained in section 2.2.

In this regard we have trained several networks in order to choice the best MLP parameters architecture using the root mean squared error RMSE as error function (Hetch-Nielsen 1990). It is the most popular measure of error and has the advantage that large errors receive much greater attention than small errors.

We note that the input and output data must be preprocessed for ANNs problems. Data preprocessing will allow the model produce accurate forecasts. Normalization and standardizing are the two most used preprocessing methods so we have tested it. Specifically we have applied a normalization scaling the data to the interval [-1, 1] and a standardization scaling inputs to have mean 0 and variance 1. The better results have been obtained with a normalization preprocessing and the hyperbolic function as activation function for the hidden and the output layer.

Finally to estimate the optimal number of hidden units we have varied the number of hidden neurons of the trained networks from 2 to 25, and we also have run several test to take account of the random initialization weights effect. We have also tested MLP with one and two hidden layers and we have obtained a better generalization with just one. The final MLP model proposed has one hidden layer with 12 neurons and the Levenberg-Marquardt (Levenberg 1944, Marquardt 1963) as training function.

Finally we add that early stopping method was used in all the run test to avoid overfitting.

3.5 *Performance of proposed ANN models*

To avoid overfitting, the ANN model should use a set different set unknown for the ANN. For that purpose, we carried out a Hold-Out validation (Bishop 1995). This method needs to divide the large dataset to three subset. Then the classifier is tested by new databases called validation and test sets. In this study the data have been randomly divided into these three subsets. The fraction of the data that is placed in the training set is 70% and 15% for the validation and test sets. However, the way the data are divide can have an important impact on model performance (Shahin et al. 2004) and the statistical properties of the various data subsets should be taken into account as a part of any data division procedure.

The matrix confusion summarize the classification of each dataset for MLP final model (Tab.3). This table is often used to describe the performance of a classification model. In this matrix, each column number represents the instance in target or actual class while each number row number represents the instance of predicted class.

Table 4. Confusion matrix for MLP proposed

		Training				Validation			
Output	Class 1	75	2	0	0	17	1	0	0
	Class 2	1	18	0	0	0	2	0	0
	Class 3	0	0	25	0	0	0	7	0
	Class 4	0	0	0	31	0	0	0	6
		Test				Total			
Output	Class 1	18	0	0	0	110	3	0	0
	Class 2	0	3	0	0	1	23	0	0
	Class 3	0	0	5	0	0	0	37	0
	Class 4	1	0	1	5	1	0	1	42
		Target class				Target class			

We note that MLP proposed has assigned the right class to 212 samples from a total of 218 that means an accuracy of 97% from the total database. The 97% and 94% accuracy have achieved for the validation and test sets respectively.

4 DISCUSSIONS AND CONCLUSIONS

In this paper we have proposed an automatic methodology to predict grain size class from dynamic penetration test and the classification task is carried out by 2 different ANN, an MLP and a PNN. The both network have achieved excellent results. Nevertheless further studies and samples may be needed to assess its application in situ.

Regarding performance of each model of ANN used in this study, we think that the 2 topologies have proved their efficiency. Nevertheless, PNN may be more interesting than MLP to detect novel cases since it is a model based on probability density estimation.

It can be noted that the application of ANNs have been possible due to characteristics of cone resistance log provided by Panda test. In addition, incorporation of other soil measurements as soil images by means of geo-endoscopy technique (Breul 1999, Haddani 2004) or Panda3 measures (Benz 2009, Escobar 2014) could be helpful. Work in this area is ongoing.

5 REFERENCES

Benz-Navarrete M.A. 2009. Mesures dynamiques lors du battage du pénétromètre PANDA 2. *Chemical and Process Engineering. Université Blaise Pascal-Clermont-Ferrand II.*

Betancourt G., Giraldo E., Franco J. 2004. Reconocimiento de patrones de movimiento a partir de señales electromiográficas. *Scientia et Técnica* vol.26: 53-58.

Breul P. 1999. Caractérisation endoscopique des milieux granulaires couplée à l'essai de pénétration. *Thèse de doctorat Physique Clermont-Ferrand 2.*

Chaigneau L., Bacconnet C., Gourvès R. 2000. Penetration test coupled with geotechnical classification for compacting control. *Proceedings of International Conference on Geotechnical & Geological Engineering, GeoEng2000 (Melbourne, Australia.*

Escobar E.J. (2015). Mise au point et exploitation d'une nouvelle technique pour la reconnaissance des sols : le Panda 3. *Thèse de doctorat Génie Civil Clemont-Ferrand.*

Fayyad U., Piattetsky-Shapiro G., Smyth P. 1996. The KDD process for extracting useful knowledge from volumes of data. *Commun ACM* 39 vol.11: 27-34.

Hadanni Y. 2004. Caractérisation et classification des milieux granulaire par géondoscopie. *Thèse de doctorat Génie Physique Clermont-Ferrand 2.*

Hayes M.H. 1996. *Statistical Digital Signal Processing and Modeling.* Ed Wiley.

Hecht-Nielsen, R. 1990. *Neurocomputing.* Addison-Wesely Publishing Company, Reading, MA.

Hornik K. and Stinchcombe M. and Halbert W. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* vol.2 (5): 359-366.

Kannatey-Asibu E., Dornfeld D.A. 1982. *Wear* vol.76: 247-261.

Levenberg K. 1944. A Method for the Solution of Certain Problems in Least Squares. *Quart. Appl.Math.* vol.1: 431-441.

Leverington D. 2009. A Basic Introduction to Feedforward BackPropagation Neural Network. *Geosciences Department Texas Tech University.*

Marquardt D. 1963. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM J. Appl. Math.* vol.2: 164-168.

Rabitz H. 1989. Systems analysis at the molecular scale. *Science* vol.246 (4927): 221-6.

Romo H.A., Realpe J.C., Jojoa P.E. 2007. Análisis de Señales EMG Superficiales y su Aplicación en Control de Prótesis de Mano. *Revista Avances en Sistemas e Informática* vol.4 (1): 128-136.

Rumelhart D.E., Hinton G.E., Williams R.J. 1986. Learning internal representations by error propagation. Parallel distributed processing: *Explorations in the microstructure cognition* vol.1: 318-362.

Saltelli A., Ratto M., Tarantola S., Campolongo F, European Commision, Joint Research Centre of Ispra(I) 2006. Sensitivity analysis practices: strategies for model-based inference. *Reliability Engineering & System Safety* vol.91 (10-11): 1109-1125.

Sarle W.S. 2002. The Neural Networks FAQ. *ftp://ftp.sas.com/pub/neural/FAQ.html*

SETRA & LCP 1992. *Technical Guidelines on Embankement and Capping Layers Construction.* Guide technique.

Shahin M.A., Jaksa M.B., Maier H.R. 2001. Artificial Neural Network applications in geotechnical engineering. *Australian Geomechanics-March.*

Shahin M.A., Jaksa M.B., Maier H.R. 2004. Data Division for Developing Neural Networks Applied to Geotechnical Engineering. *Journal of Computing in Civil Engineering* vol.18 (2), 105-114.

Shahin M.A., Jaksa M.B., Maier H.R. 2008. State of the art of artificial neural networks in geotechnical engineering. *Electronic Journal of Geotechnical Engineering.*

Shahin M.A., Jaksa M.B., Maier H.R. 2009. Recent Advances and Future Challenges for Artificial Neural Systems in Geotechnical Engineering Applications. *Advances in Artificial Neural Systems* vol. 2009: 9.

Shahrour I., Gourvès R. 2005. *Reconnaissance des terrains in-situ.* Hermès-Lavoisier.

Shannon C.E. 1948a. A mathematical theory of communication Part I. *Bell System Technical Journal* vol.27: 379-423.

Shannon C.E. 1948b. A mathematical theory of communication Part II. *Bell System Technical Journal* vol.27: 623-656.

Spetch D.F. 1990. Probabilistic neural network. *Neural Networks* vol.3: 109-118.

Sulewska M.J. 2011. Applying Artificial Neural Networks for analysis of geotechnical problems. *Computer Assisted Mechanics and Engineering Sciences* vol.18 (4): 231-241.

Waszczyszyn Z. 2011. Artificial Neural Networks in civiel engineering: another five years of research in Poland. *Computer Assisted Mechanics and Engineering Sciences* vol.18: 131-146.