American-Eurasian Journal of Sustainable Agriculture, 5(2): 168-176, 2011
ISSN 1995-0748

ORIGINAL ARTICLES

Crop Growth Prediction Using Self-organizing Map and Multilayer Feed-forward Neural Network

Ogwueleka, Francisca Nonyelum

Department of Computer Science, University of Abuja - Nigeria.

Ogwueleka, Francisca Nonyelum: Crop Growth Prediction Using Self-organizing Map and Multilayer Feed-forward Neural Network

ABSTRACT

This study outlined the extensive design of using artificial neural network for prediction of suitable crops to soil types. Preprocessing procedure was performed over the large collected data set to enhance the exactness of the mining process, which transformed the raw data adapted for clustering process. The collected soil was categorized by its geology using the self-organizing map (SOM) clustering, which helped to predict the accurate crops suitable for different kinds of soil. Six attributes were analyzed in this study. Multilayer feed-forward neural network was applied to find the best match of crops for the given type of soil characteristics after the data set was clustered. Artificial neural network technique was used because it demonstrated tolerance to a considerable degree over errors present in the training set and it also mapped out the obscured and persistently non-linear dependencies accurately.

Key words: Self-organizing map, clustering, multilayer feed-forward neural network, soil, crops.

Introduction

Agriculture is the production of goods through domesticating plants, animals and other life forms. Apart from providing food, agriculture also supplies raw materials for manufacturing industries like textiles, sugar, vegetable oils, jute and tobacco. Food and fiber production significantly depends on soil resources. Plant growth and yield measurements primarily depend on soil characteristics. Quantifying the soil characteristics with their location and other vital data pertaining to that can extract valuable information about the complex process.

The researches already done on agricultural aspect using ANN are Reddy and Ankaiah (2005). They proposed the framework for a cost-effective agricultural information dissemination system (AgrIDS), to disseminate expert agricultural knowledge to the farming community in order to improve crop productivity and they made an effort to present a solution to bridge the information gap by exploiting advances in information technology. Mai et al. (2006) demonstrated the scope for application of spatial mining tools for a utility study and analysis. The specific application of Polyanalyst gave a clear scope for evaluation and comparison of predicted and real values.

Soil is the naturally occurring, unconsolidated or loose covering on the Earth's surface. Soil is composed of particles of broken rock that have been altered by chemical and environmental processes including weathering and erosion. Soil is different from its parent rock(s) source(s), altered by interactions between the lithosphere, hydrosphere, atmosphere, and the biosphere (Chesworth, 2008). It is a mixture of mineral and organic constituents that are in solid, gaseous and aqueous states (Voroney, 2006). Soil particles pack loosely, forming a soil structure filled with pore spaces. These pores contain solution (liquid) and air (gas). Most soils have a density between 1 and 2 g/cm^3.

Most living things in soils, including plants, insects, bacteria and fungi, are dependent on organic matter for nutrients and energy. Soils often have varying degrees of organic compounds in different states of decomposition.

Corresponding Author: Ogwueleka, Department of Computer Science, University of Abuja-Nigeria.
E-mail: nonnyraymond@yahoo.co.uk

Many soils, including desert and rocky-gravel soils, have little or no organic matter; while soils, such as peat (histosols) that are all organic matter are infertile (Foth and Henry,1984). Soil covers much of the land on earth. It is made up of minerals (rock, sand, clay, silt), air, water, and organic (plant and animal) material. There are many different types of soil, and each one has unique characteristics, like colour, texture, structure, and mineral content. The depth of the soil also varies. The kind of soil in an area helps to determine the type of plants that can grow on them.

Soil separates are specific ranges of particle sizes. In Nigeria, the smallest particles are clay particles and are classified as having diameters of less than 0.002 mm. The next smallest particles are silt particles and have diameters between 0.002 mm and 0.05 mm. The largest particles are sand particles and are larger than 0.05 mm in diameter. Also, large sand particles can be described as coarse, intermediate as medium, and the smaller as fine. Soil textures are classified by the fractions of each soil separate (sand, silt, and clay) present in the soil. Classifications are typically named for the primary constituent particle size or a combination of the most abundant particles sizes, e.g. "sandy clay" or "silty clay." Loam is used to describe a roughly equal concentration of sand, silt, and clay, and lends to the naming of even more classifications, e.g. "clay loam" or "silt loam." In Nigeria, the twelve soil texture classifications are clay, silt, sand, loam, silty clay, sandy clay, clay loam, silt loam, sandy loam, loamy sand, silty clay loam, and sandy clay loam.

Self-organizing maps (SOMs) are a data visualization technique invented by Professor Teuvo Kohonen, which reduce the dimensions of data through the use of self-organizing neural networks (Germano, 1999). Kohonen's SOMs are a type of unsupervised learning (Kohonen, 1997). The first part of a SOM is the data. The second component to SOM is the weight vectors. Each weight vector has two components to them. The first part of a weight vector is its data (Kohonen, 1988). This is of the same dimensions as the sample vectors and the second part of a weight vector is its natural location. Kohonen's SOM is called a topology-preserving map because there is a topological structure imposed on the nodes in the network. A topological map is simply a mapping that preserves neighborhood relations. It considers the physical arrangement of the nodes as nodes that are "close" together are going to interact differently than nodes that are "far" apart. The aim is to train the net so that nearby outputs correspond to nearby inputs. In the brain, neurons tend to cluster in groups.

A multilayer feed-forward network was used because of its genuine characteristics. This study was to find a solution that could provide precise information on the suitability of crop(s) for a given soil type. The dataset analyzed was diversely collected from agricultural experts pertaining to the area under research. The dataset are the collected data, which are presented in tabular form under results. It has columns and rows, where each column represents a particular variable while each row corresponds to a given member of the dataset. The dataset has several characteristics, which define its structure and properties, including the number and types of the attributes. The seven attributes used are elevation, slope, erosion, drainage, permeability, geology, and taxonomic classification while the parameters used are land use and vegetation. With the object of increasing the accuracy of mining process, sufficient preprocessing was performed over the collected dataset, which transforms the raw data for the clustering process.

Methodology:

Multi-Layer Perceptron Neural Network (MLPNN) also known as multilayer feed-forward neural network was chosen and used in this study. The multilayer feed-forward neural network used back-propagation learning technique. The soil geology was categorized using SOM based clustering. This process of stratification goes up till a level enough to predict near accurate crop(s) suitable for that particular kind of soil. The attributes analyzed include the elevation, slope, erosion, drainage, permeability, geology, and taxonomic classification. The parameters used are land use and vegetation. Land use show the crops tested for that particular type of soil; and vegetation signifies the natural growth of crops in that region. Some horizon attributes like colour, texture, type, and amount of pores were measured. Figure 1 is the multilayer feed-forward neural network showing the attributes and parameters.

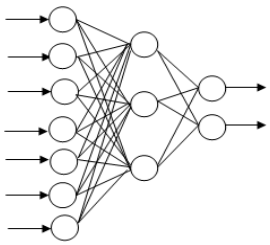

Fig. 1: Multilayer feed-forward neural network.

This approach of stratification and clustering was used to find a solution that could provide specific information on the fitness of crop(s) for a given soil type. After the complete datasets are clustered, the neural network was applied to find the best match of crop(s) for the given type of soil characteristics as NN learn from examples once the training is completed and so makes it an appropriate methodology for prediction. The NN was trained to find the best match of soil and its suitable crop(s) through thorough analysis of the characteristics of the soil and the nutritional requirements of the crop(s) planned for cultivation. This was done through a database built through observation of crop(s) cultivated previously in the soil types, its growth patterns, and laboratory experiments. The output of this process was the list of cultivable crops and crops that could grow naturally. This ANN system leads to minimal application of fertilizers and appropriate irrigation patterns.

The NN exhibited tolerance to a substantial degree over errors that maybe present in the training set. Its capability to map out the veiled and non-linear dependencies, and still learn from examples once the training is completed makes it an appropriate methodology for prediction. Figure 2 and 3 Ogwueleka and Ogwueleka, 2009)shows the overall architecture of the designed system with the neural network model.

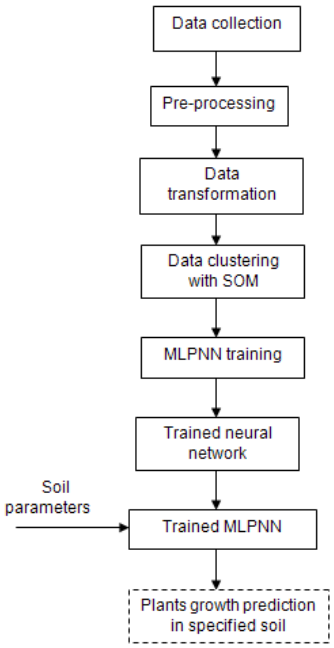

Fig. 2: The training architecture.

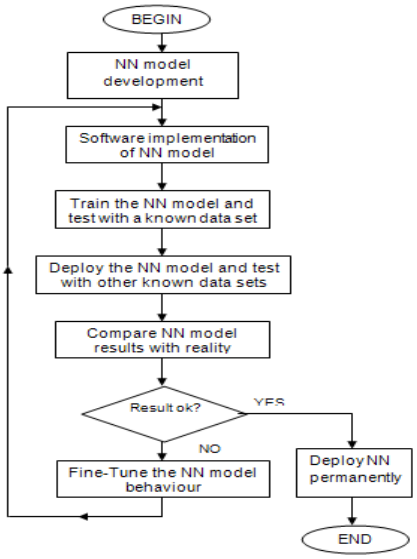

Fig. 3: Steps in ANN model design.

The data in the databases are collected from local experts. The data are preprocessed through duplicate removal and supplying of missing values. Data was transformed to a format suitable for clustering because data elements with high similitude are bunched up within same cluster while data elements with less similarity are crowded up in different clusters through unsupervised classification of patterns into groups. To prepare the data for mining process, it needs to be transformed accordingly. This preprocessed data is refined further and converted to a format compatible for clustering. Clustering was performed with k-means algorithm. SOM was used as the basis so clustering of data item was an approach employed. The methodology projected takes advantage of k-means measurement of distance among the sampled data because it was the main constituent of the cluster analysis. When the entire datasets are clustered, using self-organizing map, then the ANNs are applied to find the optimum match of crop(s) for the specified type of soil characteristics. The ANN was trained to find the best type of soil and its suitable crop(s) through detailed analysis of the soil characteristics and the nutritional requirements of the crop(s) planned for cultivation. The prediction was done using Multilayer Perceptron Neural Networks (MLPNN).

One of the simple algorithms used in this research is:
Initialize map
For r from 0 to 1(where r is the weight vector)
Randomly select a sample
Get best matching unit
Scale neighbours
Increase r with small amount
End for

The first step in constructing this SOMwas the initialization of the weight vectors. A sample vector was selected randomly and the map of weight vectors searched to find which weight represented that sample excellently. Since each weight vector has a location, it also has neighbouring weights that are close to it. The weight that was chosen was rewarded by being able to become more like that randomly selected sample vector. Also, the neighbours of that weight are also rewarded by being able to become more like the chosen sample vector. From this step, r was increased with small amount because of the number of neighbours and how much each weight can learn to decrease over time. This whole process was repeated a large number of times.

The goal of the training process of the multilayer feed-forward network is to find the set of weight values that will cause the output from the neural network to match the actual target values as closely as possible. There are several issues involved in the proposed designing and training of the multilayer perceptron network, such as, selecting how many hidden layers to use in the network; deciding how many neurons to use in each hidden layer; finding a global optimal solution that avoids local minima; converging to an optimal solution in a reasonable period of time; and validating the neural network to test for overfitting.

Despite the presence of considerable noise and errors in the training set, ANN has the ability to trace the hidden and strongly non-linear dependencies, and learn from examples once the training is completed. It is an appropriate tool when it comes to prediction.

Results:

The designed MLPNN consist of one input layer, two hidden layers and one output layer. The output layer predicts the crops growth in a soil given by its characteristics. Table 1 shows some of the soil characteristics and weights assigned to it for MLPNN training. In Table 2, the input soil characteristics for predicting the MLPNN is shown, while Table 3 displays the output weights of the MLPNN training for land use and natural vegetation.

Table 1(A-F): Weights of the parameters in the MLPNN training.

A. Elevation

Elevation	Training weight
1 m above MSL	0.01
100 m above MSL	0.0100
101 m above MSL	0.0101
102 m above MSL	0.0102
103 m above MSL	0.0103

Table 1: Continue

B. Slope	
Slope	Training weight
Flat lands	0.01
Foot and side slopes	0.02
Foot slopes	0.03
Mild slope with depression	0.04
Mild slope	0.05
C. Erosion	
Erosion	Training weight
Moderate erosion	0.01
Moderate to severe erosion	0.02
Severe erosion	0.03
Gully erosion	0.04
Very severe erosion	0.05
D. Geology.	
Geology	Training weight
Clay	0.01
Granite	0.02
Laterite	0.03
Quartzite	0.04
Sand	0.05
E. Permeability.	
Permeability	Training weight
Medium permeability	0.01
Moderate permeability	0.02
Moderate to slow permeability	0.03
Moderately slow permeability	0.04
Rapid permeability	0.05
F. Drainage.	
Drainage	Training weight
Excessively drained	0.01
Imperfectly drained	0.02
Moderately drained	0.03
Poorly drained	0.04
Poorly dained	0.05

Table 2: Input soil uniqueness for predicting the MLPNN.

Elevation	Slope	Erosion	Geology	Permeability	Drainage
100m above MSL	Mild slope	Gully erosion	Granite	Moderate permeability	Poorly drained
1m above MSL	Flat lands	Moderate erosion	Clay	Medium permeability	Excessively drained

Table 3: Land use and natural vegetation output weights of the MLPNN training.

Land use	Training weight
Banana and guava	1
Banana and mango	2
Banana and sugar cane	3
Banana	4
Maize	5
Natural vegetation	Training weight
Acacia, wet weeds, grasses	1
Acacia, wet weeds, calotropis	2
Acacia, cassia auriculata	3
Acacia, croton sparciflorus	4
Acacia, prosopis julifora	5

Table: 4 Predicted land use and crops for the given soil type.

Land use	Natural vegetation
Banana	Acacia, cyanodon, neem
Maize	Acacia, prosopis julifora

Discussion:

Multilayer feed-forward neural network is the class of networks that consist of multiple layers of computational units, usually interconnected in a feed-forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer. Multi-layer networks use a variety of learning techniques, the most popular being back-propagation and this was used in this study. The output values are compared with the correct answer to compute the value of some predefined error-function. By various techniques, the error

is then fed back through the network. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function to the minimum. After repeating this process for sufficiently number of training cycles, the network converged to some state where the error of the calculations was small, and then the network has learned a certain target function. To adjust the weights properly, a general method was applied for non-linear optimization, called gradient descent. For this, the derivative of the error function with respect to the network weights was calculated, and the weights are then changed such that the error decreases.

In this multiplayer perception feed-forward neural network, the neurons of the first layer steered their output to the neurons of the second layer, in a unidirectional manner so that the neurons are not received from the reverse direction. Incorporating the three layers- input, output and intermediate (hidden), the MLPNN designates distinct roles for each. Input layer maintains equal number of neurons corresponding to that of the variables in the problem. The output layer comprises a number of neurons equal to the used number of quantities computed from the input and made accessible the perceptron responses. The intermediate/hidden layer takes care of approximating non-linear problems. Processing linear problems necessitates the presence of only the input and output layer of the MLPNN.

The only task of the neurons in the input layer is the distribution of the input signal to neurons in the hidden layer. Each neuron in the hidden layer sums up its input signals after weighting them with the strengths of the respective connections from the input layer and computes its output as a function of the sum. The output of neurons in the output layer was computed in the same manner. Following this calculation, a learning algorithm was used to adjust the strengths of the connections in order to allow the network to achieve the expected behaviour.

Back propagation (BP) training was used as it was considered to be the most versatile algorithm used to train the network for predicting correct outputs especially outputs obtained from experiments and generated one. The error information was propagated backward through the network. The BP algorithm adjusts the network weights and bias values to minimize the square sum of the difference between the given output and output values calculated by the net.

The backpropagation technique was used efficiently by these steps :

1. A training sample was presented to the neural network.
2. The network's output was compared to the desired output from that sample and the error in each output neuron was calculated.
3. For each neuron, what the output should have been, a scaling factor, and how much lower or higher the output must be adjusted to match the desired output was noted and the local error noted.
4. Each neuron’s weights was lowered and the local error was adjusted.
5. Fault was assigned for the local error to neurons at the previous level, giving greater responsibility to neurons connected by stronger weights.
6. Repetition from step 3 on the neurons at the previous level, using each one's fault as its error was done and the findings noted.

In this study, the conjugate gradient algorithm was used to adjust weight values using the gradient during the backward propagation of errors through the network. Compared to gradient descent, the conjugate gradient algorithm takes a more direct path to the optimal set of weight values. Usually, conjugate gradient is significantly faster and more robust than gradient descent. Conjugate gradient also does not require the user to specify learning rate and momentum parameters.

The steps used are as follows;

1. Calculate total number of items in the dataset
2. Form number of clusters and calculate the centroid of the clusters.
3. Calculate the Euclidean distance of each data item with the centroids of the available clusters.
4 .Assign the data item to the cluster with the minimum distance
5. Repeat steps 3 and 4 until there is no change in the clusters.
6. Calculate the standard deviation of all the clusters formed. Neglect clusters with standard deviation less than 0.
7. Repeat steps 1 to 6 until the standard deviation of all clusters reaches a value lless than 0.

The pseudocode of the clustering algorithm is as follows:

Initialize N_C, N_o (where $N_C = 1$ and $N_o = 1$)
Calculate N_C (where $N_C = N_C - N_o + 1$
Form C_{if} clusters of N_C size
Calculate centroid
Repeat
For values of 1 to N and N_C

Calculate Euclidean distance
Perform minimum distance
Assign data item to cluster until there is no change in formed clusters
For cluster values equal to 1 to N_C
Calculate standard deviation
If standard deviation is less than Φ
Remove C_r from C_{if} where C_r € C_{if} and add to C_a
Compute the remaining clusters until its reaches zero

where N = Total number of items in the dataset
N_C = Number of clusters
N_o = Number of clusters with zero standard deviation
C_a = Actual clusters
C_{if} = Intermediately formed clusters
C_r = Cluster to be removed.

The flow diagram of the clustering process is shown in Figure 4. The initial value of N_C = 1 and N_o = 0. After calculating the centroid and Euclidean distance of the data items, the minimum distance grouping is done.

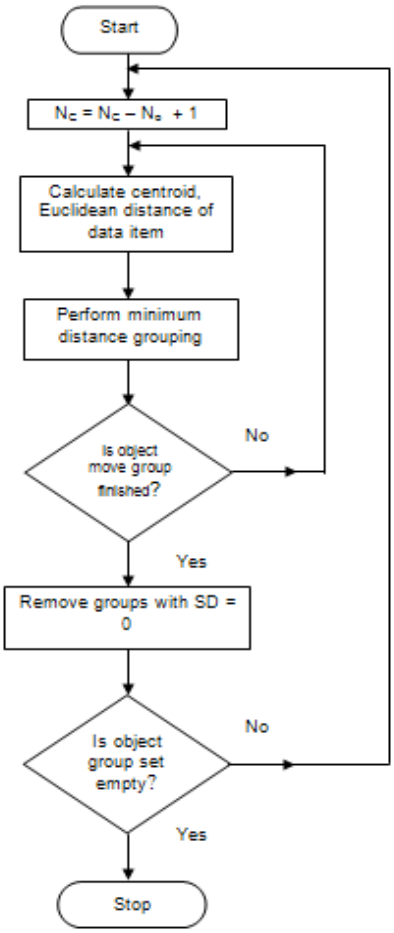

Fig. 4: Clustering process flowchart.

The functionality of the hidden neuron was mathematically described with the equation

$$\sigma\left(\sum_{j=1}^{n} w_{jx} x_j + b_j\right)$$

where the weights w_j and b_j are symbolized with the arrows feeding into the neuron. The network output was formed by another weighted summation of the outputs of the neurons in the hidden layer. This summation on the output was called the output layer. The number of output neurons equaled the number of outputs of the

approximation problem. The output of this network was obtained using-

$$\hat{y}(\theta)=(\theta,x)=\sum_{i=1}^{nb} w_1^2\,\sigma\left(w_{jx}\,x_j+b_j\right)$$

where n is the number of inputs and nh is the number of neurons in the hidden layer. The variables $(w_{i,j}^1, b_{j,i}^1, w_i^2, b^2)$ are the parameters of the network model that are represented collectively by the parameter vector. The neural network model was represented by the compact notation $g(\theta,x)$ whenever the exact structure of the neural network is not necessary in the context of a discussion.

The operation of the network was divided into two phases: the classification phase and the learning phase. In training the network, its parameters were adjusted incrementally until the training data satisfied the desired mapping as well as possible. During the learning phase, the weights in the feedforward network were modified. All weights were modified in a way that when a pattern was presented, the output unit with the correct category always has the largest output value.

The feedforward network designed used a supervised learning algorithm, so besides the input pattern, the neural network also require to know to what category the pattern belongs. Learning proceeds as follows: a pattern was presented at the inputs and transformed in its passage through the layers of the network until it reaches the output layer. The units in the output layer all belong to different categories. The outputs of the network as they are now are compared with the outputs as they ideally would have been if this pattern were correctly classified: The unit with the correct category had the largest output value and the output values of the other output units was observed to be very small. On the basis of this comparison, all the connection weights are modified a little bit to guarantee that the next time this same pattern was presented at the inputs, the value of the output unit that corresponds with the correct category is a little bit higher than it was presently and that, at the same time, the output values of all the other incorrect outputs are a little bit lower than they are presently too. The differences between the actual outputs and the idealized outputs are propagated back from the top layer to lower layers and used at these layers to modify connection weights.

Performing the procedure once for every pattern and category pair in the data set, meaning that 1 epoch of learning has been done. After many epochs, the neural net was noted to consider and remember these pattern-category pairs making it able to generalize when the neural net's learning phase was terminated. It learnt to classify correctly any unknown pattern presented to it. In this study, three different objects were selected – the feedforward net, which is the classifier; the pattern, which is the inputs; and the category, which is the correct output.

The learning phase duration depended on the size of the neural network, the number of patterns to be learned, the number of epochs, the tolerance of the minimizer, the speed of the computer, and the computing time the learning phase took. The approach was that if computing time becomes excessive in the interactive environment then the scripting facilities would be used to process the learning job as a batch.

In the classification phase the weights of the network were fixed. A pattern, presented at the inputs was transformed from layer to layer until it reached the output layer. Classification was also through the selection of the category associated with the output unit that had the largest output value. For classification, a feedforward net and a pattern are selected together while categories were chosen from.

Conclusion:

A good estimation of suitability of soil characteristics and the choice of crops planned for cultivation based on its nutrient requirements was accurately predicted through the application of SOM and MLPNN. SOM was used for clustering because of its ability to withstand noise levels to certain extents, the capacity to learn in a self-paced manner and the degree to which it can self-govern. The MLPNN was also used for predicting the clustered datasets because of its ability to accommodate multiple intermediate layers.

The results obtained proved that NNs are the best for prediction of crops growth in different soil types.

Out of the several techniques of artificial intelligence like Expert Systems, Fuzzy Logic, Genetic Algorithm, Intelligent Agent, Neural Networks, Neural Networks was used because it exhibits tolerance to a substantial degree over noises if present in the training set and learn from examples once the training is completed, making it an apt methodology for prediction.

References

Chesworth, 2008. Encyclopedia of soil science, Dordrecht, Netherland: Springer, xxiv.

Demiriz, A., K.P. Bennett and M.J. Embrechts., 1999. "Semi-supervised clustering using genetic algorithms," In Artificial Neural Networks in Engineering (ANNIE-99), pp: 809-814.

Foth and D. Henry, 1984. Fundamentals of soil science, New York: Wiley, pp: 151.

Kohonen, T., 1998. Self-Organization and Associative Memory,New York: Springer-Verlag.

Kohonen, T., 1997. Self-Organizating Maps, New York : Springer-Verlag.

Mai, C.K., Murali Krishna and A.V. Reddy., 2006. Data Mining Of Geospatial Database for Agriculture Related Application. Map India.

Ogwueleka, T.C. and F.N. Ogwueleka, 2009. Application of artificial neural networks in estimating wastewater flows. The Icfai University Journal of Science and Technology, 5(3): 20-30.

Reddy, P.K. and R. Ankaiah., 2005. "A framework of information technology-based agriculture information dissemination system to improve crop productivity". Curr. Sci., 88: 1905-1913.

Voroney, R.P., 2006. The Soil Habitat *in* Soil Microbiology, Ecology and Biochemistry, Eldor A. Paul ed. ISBN=0125468075.