IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 6, June 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

Statistical Analysis of Engineering Properties of Alluvial Deposits in Western Mashhad City

Mahin Etemadifar[1], Nargess Alizadeh Loushabi[2], Iman Aghamolaie*[3], Saeedeh Hossein[4]

[1] Masters of Geological Engineering Abpouy consulting engineering company
[2]Masters of tectonic geology, Islamic Azad university of shahrood
[3]Masters of Geological Engineering
[3] Lecturer, Department of Geology, Payame Noor University, PC 19395-3697,Tehran, Iran

Abstract

West of Mashhad is made of alluvial fan or cone-shaped deposit and is built up by sand and gravel with little silty interlayers to a depth of 30m. In the present study, data of more than 190 boreholes were collected and stored in an appropriate database. Based on the statistical analysis, main properties of the deposits including cohesion, internal friction angle, standard penetration test, plastic limit, liquid limit, soil material and grain size distribution were obtained. For achieving a classification with a suitable precision, soils were compared with each other by introducing a huge number of training classes. In this research, k-mean clustering analysis method was used for soil classification and for providing similar physical characteristics of each soil class. Based on this method, soils in two different regions in terms of depth (0-5 and 5-10 m) are clustered in 4 different clusters with the highest similarity degree to each other and the relations between different parameters are obtained.

Keywords: *Geotechnical database, clustering, soil texture, Mashhad.*

1. Introduction

Spatial interpolation was used in order to generalize the geotechnical data. In fact in this method, existing data are rebuilt based on adjacent region conditions. There are different algorithms for spatial interpolation. Generally theory of regional variables is used in geology which considers both structural and eventual variations of spatial variables. Despite all developments which are made in spatial modeling, sufficient attention should be paid for the results obtained from these models. One of the main tools of geostatistics for investigation of spatial variations and parameter properties is semi-variogram. In This method, characteristics of two values of Z(x) and Z(x+h) which are coordinated at x and x+h and in a distance of h from each other are investigated. Theoretically, the value of semi-variogram for h=0 should be reduced to zero. But practically real half variation of the exponents are not equal to zero because of varying nature of depositional environment, sampling error , data preparation and measurement and analysis which is called piecewise error. The studied region in western Mashhad has a longitudinal coordination of 59^{o} $27^{'}$ to 59^{o} $33^{'}$ west and 36^{o} $18^{'}$ to 36^{o} $23^{'}$ north (Yusefi, 2003).

2. Research Method and Discussion

A Database is an organized set of data. In a database, a set of records are stored in a computer, by a systematic method and principals like a computational program and can be retrieved by the user. A computational program which controls the storage and restoration of organized data is called a database management system or in abbreviation DBMS. This system controls database security and robustness (Maliki, 2002: Suwanwiwattana et al, 2001: Swift, 2002).
Database design is a process of decision making about the method of data organizing in different rows and establishing the relation between the rows. Figure 1 shows the data base design process (Yongli Gao et al, 2002: Lugo, 2007: Luna et al, 2001: May, 1999).

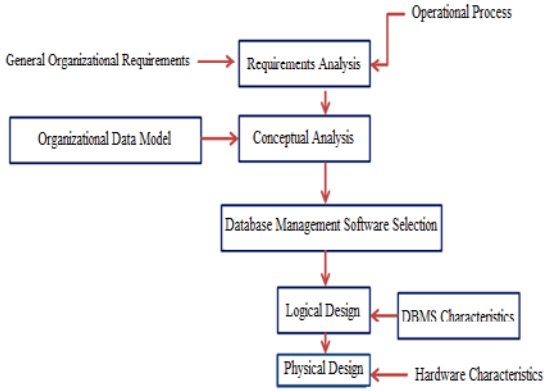


Figure 1. Database design process

In order to build the database of Mashhad city, different geotechnical reports were used which were prepared by governmental and private companies. In this regard, data of up to 191boreholes were collected.
Maps existing in this database were prepared in Arc View GIS software and were connected to this database. Database tables were designed in Microsoft Access 2010. Figure 2 shows a screenshot of the worksheet form.

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 6, June 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

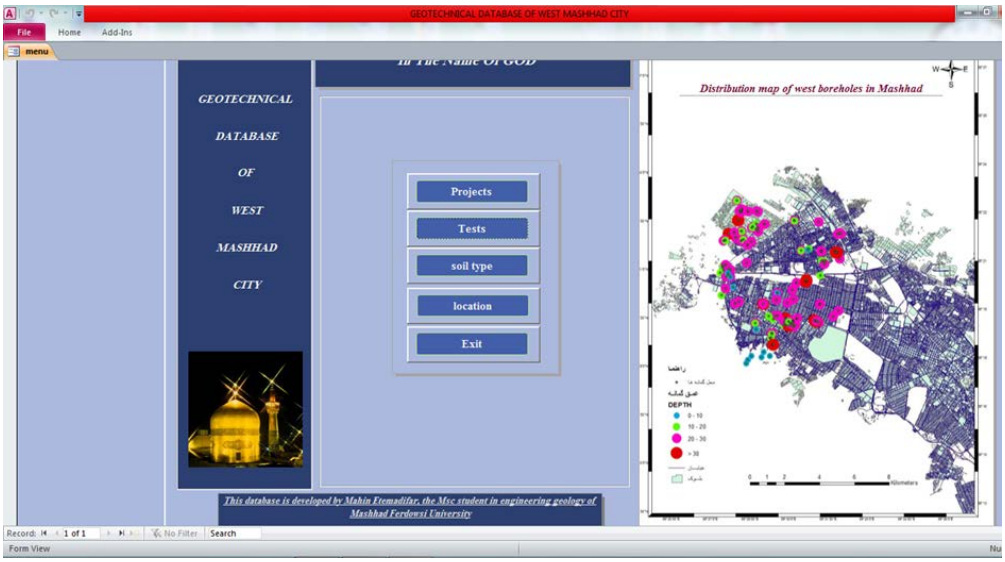


Figure 2. A screenshot of worksheet form

3. Structure design and information organizing

Determining the data structure for database is the most important part of work. There are 8 tables in the western Mashhad database. Data access is possible in 4 ways: Projects, tests, soil type and borehole location on the map. Figure 3 shows the structure and relations between tables in the western Mashhad geotechnical database.
Content of tables are as follows:
Project: includes project characteristics, operation location and project advisor.
Borehole: includes borehole number, final depth of borehole and final coordination of borehole.
Soil type: includes different types of existing soils, depth and location of the soil type.

Mechanical, physical and chemical characteristics: includes physical characteristics such as dry and wet special weight, moisture percentage, Atterberg limits in different depths and remaining percentage , mechanical characteristics such as direct shear and triaxial shear test and standard penetration test and chemical characteristics such as chloride and sulfate percentage and alkalinity (pH) of the soil. Figure 3 shows the structure and relations between tables in western Mashhad geotechnical database.

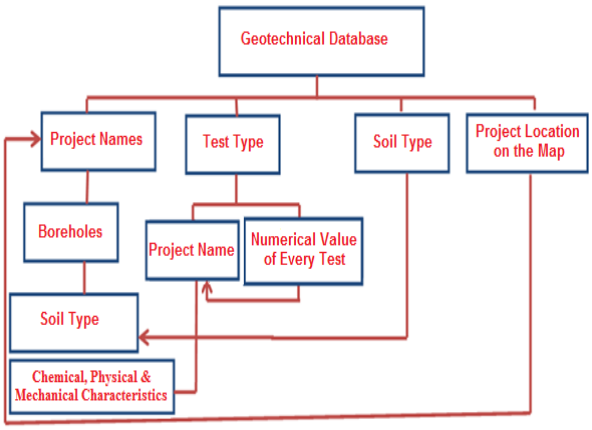


Figure 3. Shows a data entry form.

4. Soil texture clustering

The term cluster analysis was used by Tryon in 1939 for classification method for similar objects. Cluster analysis is a shortcut tool for analyzing data which aims to arrange different objects into groups which relation degree between two objects is maximum if they belong to one group and otherwise it is minimum. In other words cluster analysis shows the data structure without explaining what exists. Of course this clustering is performed when dispercial of a society from which samples are taken is very high. Meanwhile the main condition for creating clusters is that clusters are members from the society or the sample. It means that every data would be placed just in one cluster and on the other hand sum of all the clusters equals the whole considered sample or the society. Hence the purpose of clustering is to classify observations into

homogeneous groups so that observations of every group will have the most similarity and observations of different groups will have least similarity to each other. In clustering, concepts of distance and similarity are one of the fundamental concepts. Distance is a measure that shows to what extent two observations are apart from each other. On the other hand similarity is a measure of their closeness to each other. For this reason prior the analysis, it is necessary to select a quantity scale which is measured based on similarity among observations. These criterions are selected by considering clustering construction algorithm, variable nature (continuous, discrete or two-valued) and measurement scale.

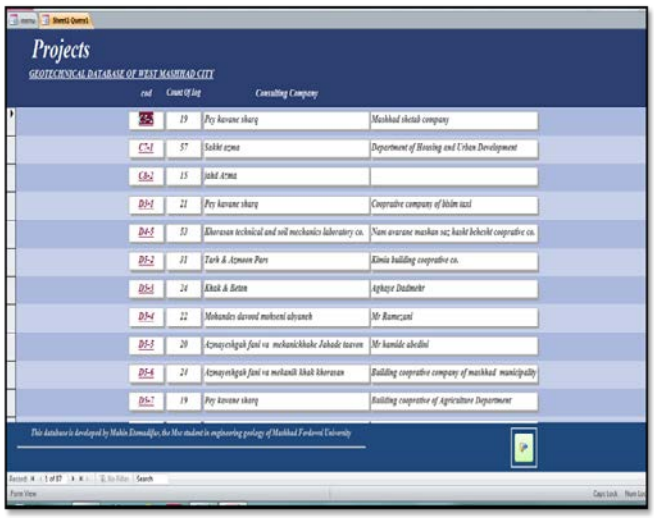


Figure 4. Project form

In the present study, statistical analysis is performed on 3600 data related to 191 boreholes in the west region of Mashhad using SPSS18 and MATLAB. For this reason, first data were normalized in order to prevent from the influence of data transmittal variance. In this investigation, subjective analysis is used for reducing data volume and determining the most effective variables in the formation of phenomena. This method was introduced by Pierson and Spearman in 1986 for the first time while measuring artificial intelligence and it is used for determining the most effective variables when the number of variables under investigation is high and the relation between them is unknown. Based on this method and in this research, variables are placed in agents in such a way that from the first agent to the subsequent agents, variance percent decreases. In this manner variables which are placed higher are most effective. For selecting the appropriate variables for the objective analysis, correlation matrix was used in this research. And for making sure about the appropriateness of data for the objective analysis besides assurance about that correlation matrices are not equal to zero which are basis of the analysis in the society, we used Croit-Bartlet test based on equation 1.

$$X^2 = -\left(n - 1 - \frac{2p+5}{6}\right)\ln|R| \qquad (1$$

In which n represents the number of subjects and p is the absolute determinate of correlation matrix. It's obvious that the statistics has a cubic distribution X (X^2) with a freedom degree of $0.5p(p-1)$.

In order to correct the standard penetration number, the suggested equation of Liao and Whitman is used. In this method, the necessary adjustment coefficient for the correction of standard penetration number is obtained from the equation 2 (Liao & Whitman, 1986).

$$C_N = \sqrt{\frac{1}{\sigma}} \qquad (2$$

In which σ is the effective overload stress in ton/ft^2. Equation 2 converts into equation 3 in SI system.

$$C_N = 9.78\sqrt{\frac{1}{\sigma}} \qquad (3$$

In which σ is the effective overload stress in KN/m^2. With respect to this equation, as the effective overload stress increases, standard penetration number reduces by reverse root.

4.1. Hierarchical clustering algorithm

Hierarchical clustering has this ability to determine the number of appropriate groups without any assumptions and can perform this action while several variables exist for the classification. In such a condition that numerous variables exist in the classification or the classification is going to be performed in a multidimensional space and almost there's no information about data distribution, this method can be the best choice for data classification (Vahaaho et al, 2003). In this clustering method, since the scoring function attributes a score to each observation membership in the cluster therefore for any new observations, the membership in one of the clusters can be determined with respect to the score which is obtained for each membership in the clusters. We attribute the observation to a cluster which the new observation attribution's score to a group with the higher score. Stages of cluster construction algorithm are shown in table 1. In this algorithm, entries, used method, measurement method, data normalization and minimization, maximization and increments of clustering can be observed clearly.

Table 1. Algorithm of a cluster construction

PROXIMITIES SPT C Phi Ms. Dd PP LL PI
/MATRIX OUT ('C: \clus.tmp')
/VIEW=VARIABLE
/MEASURE=SEUCLID

/STANDARDIZE=NONE*
/STANDARDIZE=VARIABLE RESCALE **
CLUSTER
/MATRIX IN('C: \clus.tmp')
/METHOD ={BAVERAGE} [Linkage]
/PLOT =[VICICLE DENDROGRAM [(0[8[1]])]
/PRINT =[CLUSTER({0,8})][1]]
(* For the first cluster, ** For the second cluster)

With respect to great differences between numerical values of data it's necessary to normalize the data first. For this reason data are divided by their maximum value and the result will be between zero and one. It will lead to erroneous results in the hierarchical analysis if we do not perform this operation. Figure 5-a shows the results of hierarchical clustering analysis by intergroup mean continuity method without standardization and figure 5b shows the results after standardization. It can be observed that in the first case, the standard penetration number shows a very week relationship with other parameters while in the second case it shows a logical relationship.
In figure 5, it can be observed that there is a meaningful relationship between the variables. For example, the amount of cohesion and the existing clay volume in the soil sample, liquid limit and the existing moisture content in the soil, internal friction angle and the amount of density (which is itself correlated with the soil type) and the standard penetration number and soil plasticity index have the minimum distance to each other. This is a relationship which cannot be seen in figure 5 –a.

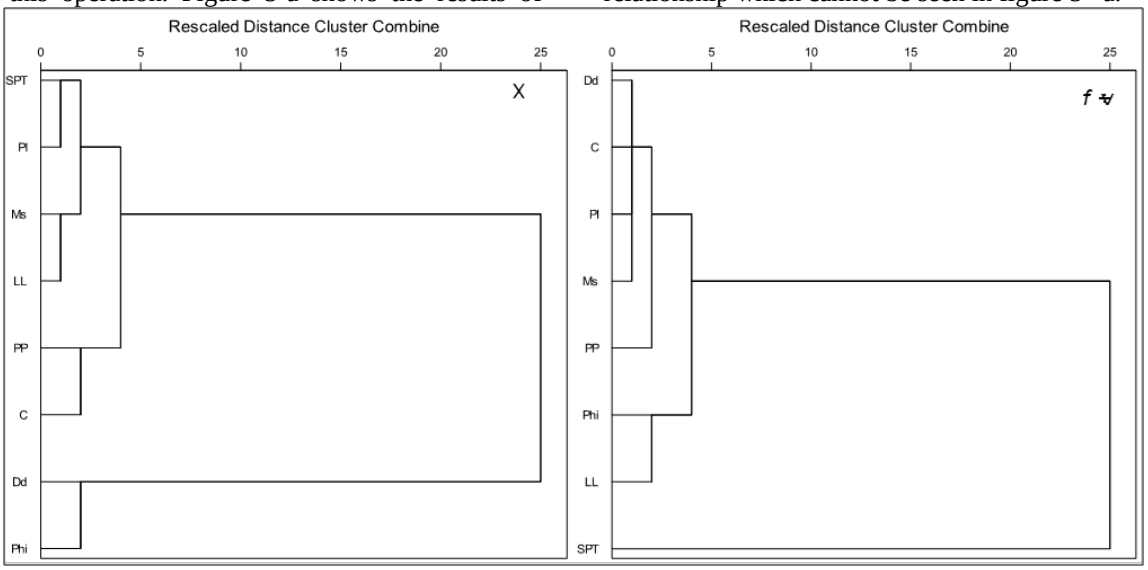
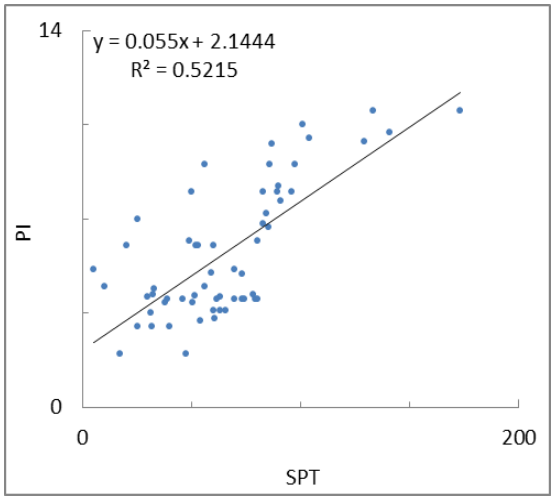


Figure 5. Dendro grams drawn by intergroup mean continuity method for tested variables in a depth from 0 to 5 m (a: before the data standardization b: after the data standardization).

4.2. Linear regression

Existence or lack of a relationship between two variables can be shown in a plot called scatter plot. In the present study there is a linear relationship between parameters which show a meaningful relationship in cluster analysis. Results are shown in figure 6 to 8.



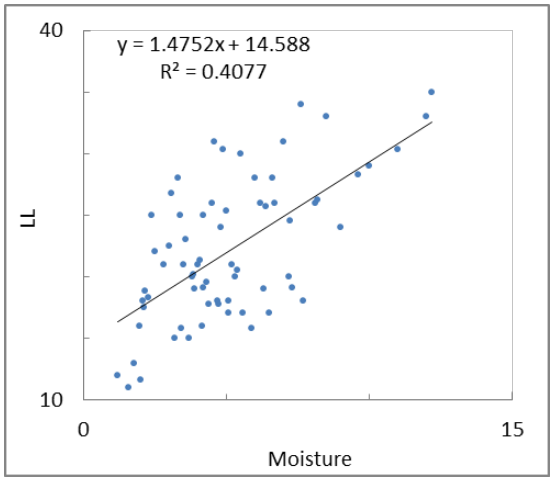
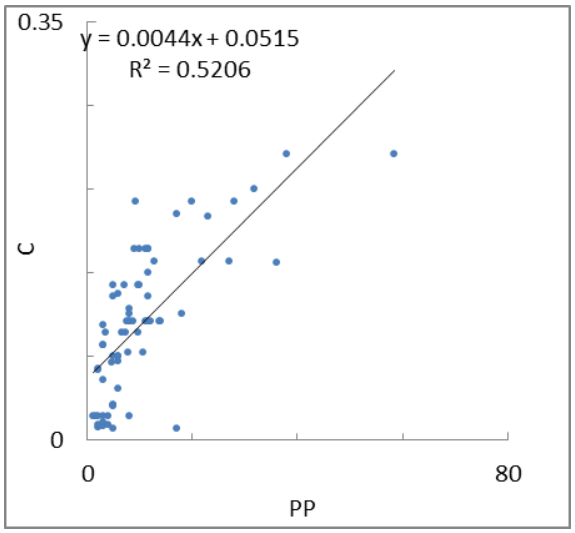


Figure 6. Linear relationship between PI and SPT, LL and Moisture



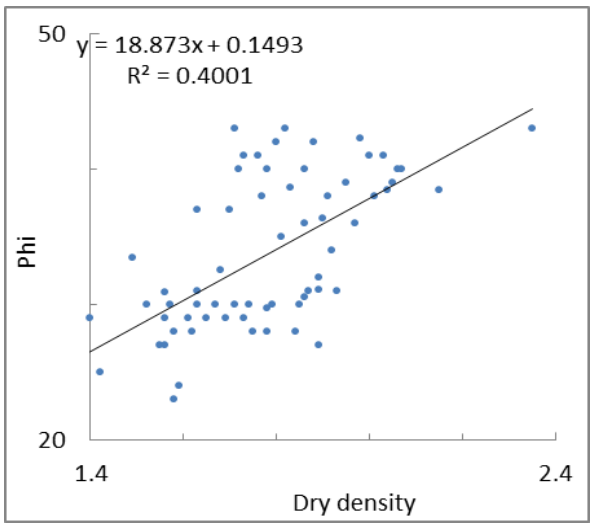


Figure 7. Figure 6. Linear relationship between C and PP, Phi and Dry density

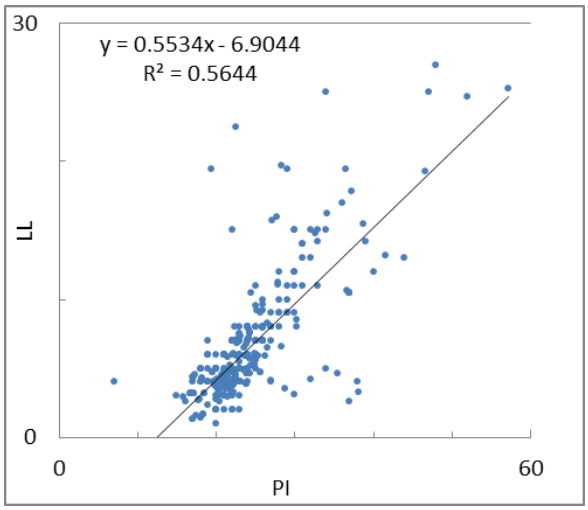


Figure 8. Linear relationship between PI and LL

4.3. K-means clustering algorithm

Linear regression results of all data sets show that variables under investigation have a meaningful relationship with each other. But presence of all data besides each other causes R^2 to decrease. For example in this analytical method of regression analysis, fine grained soil samples which have naturally a lesser SPT range in compare with coarse grained soils cause error to increase. In order to solve this problem, prior the linear regression it's necessary to perform a kind of data separation to enhance precision of the results. With respect to the impact of fine grained soils on soil characteristics, western Mashhad soil samples were separated on the basis of fine grain percentage. For this reason, existing geotechnical data matrix were read in MATLAB software, the matrix were prepared and entered in K mean algorithm. Parameter separation and data clustering were done, considering how closed is the Euclidean distance between the data and cluster center to each other. Ultimately 4 clusters were constructed for the soils of the region.

It is worth mentioning that K mean algorithm is a learning algorithm without supervisor for input data classification. In this method, points are placed in different classes with respect to their inherent distance from each other. In this algorithm it is supposed that data compose a vector space and we attempt to perform clustering operation on them. Data are clustered around the core centers which are obtained by minimizing objects. Equations 4 and 5.

$$\mu_j \forall i = k \dots 1 \quad (4$$

$$V = \sum_{i=1}^{k} \sum_{xj \in s_i} (x_j - y_i)^2 \quad (5$$

In which k is number of clusters (S_i , i = 1, 2, … , k) and μ_i is cluster center or middle points $X_j \in S_i$.

In continue, clustering analyses results are presented for data sets.

Cluster 1: coarse grained soils (sand and gravel) with remarkable amount of silt and clay and bad grained gravel and sand are main composers of soils of this group. Lowest values of SPT test for every decrease in depth is generally seen in this cluster (values lower than 40). Its cause can be attributed to the presence of high level of muck and clay in the soil of this cluster and since SPT values are generally lower than 40 therefore soils in this cluster have a semi-compacted status. The Lowest value of dry density is generally seen in this cluster (values lesser than 1.9). This is because of loose to semi-compacted soils in this cluster. Relative density of soils can be reported in the range of 0.15 to 0.65 by approximation. Among the soils of this cluster, we can refer to GP, SP, GP-GC, SP-SC and SP-SM.

Cluster 2: coarse grained soils (sand and gravel) with little amount of silt and clay and well grained sand and gravel are Main characteristics of the soil samples in this group. Highest value of SPT test for every increase in depth is generally seen in this cluster (values more than 25). Soil samples of this cluster can be called as materials without cohesion. Relative density value for the soils of this cluster can be reported in around 0.8. Therefore this group's soils have high dry densities. Well integration of the porous net and small amount of inter-grain cement in the soils of this cluster can create a very high permeability in the samples. Among the soils of this group we can refer to GW, SW, GW-GM and SW-SM.

Cluster 3: the amount of clay and muck in the soils of this cluster increases in compare with the two previous clusters. This causes a little inconsistency in the physical characteristics of the soils. For example, soils of this cluster have a high value of cohesion and average amount of internal friction angle which can be attributed to the presence of fine and coarse particles in the soils. Therefore it is likely that clay and muck with low plasticity characteristics to be inclined to sand and gravel in sense of physical characteristics like internal friction angle and SPT values. This case sounds possible if only this type of soils is compacted. And also liquidity limit and plasticity index for soils of this cluster are more close to CL-ML soils. Among the soils of this cluster we can refer to SC-SM and GC-GM.

Cluster 4: soils of this cluster are very similar to soils of cluster 3 and the only difference is the uniformity intensity of soil. Uniformity causes soils of this cluster to have a higher amount of SPT and internal friction angle in compare with the previous clusters. Among the soils of this cluster we can refer to SC, SM, GC and GM. In figure 9, soil texture map is drawn based on built clusters in this study. As you can see in this figure, soil texture can generally be divided into 4 groups. In the northern part we can see soil type 1, in the middle part, we can see soil type 2 and in the southern part, soils related to cluster type 4 can be seen. Soils related to cluster 3 has a limited expansion to northwest region.

This map is drawn by ArcGIS 9.3 Software by interpolation. As you can see in the map, in the northern part of the area of study, soil type 1 is dominant and in the middle part, soil type 2 and in the southern part, soil type 4 are dominant. Soil type 3 has a limited expansion to northwest.

K mean Clustering results are shown in figure 10. It's worth mentioning that in this figure, clusters which have nearly the same results are merged with each other. And also clusters which lack a meaningful value are not shown in the figure. Results infer that coefficient of R2 decreases after performing clustering.

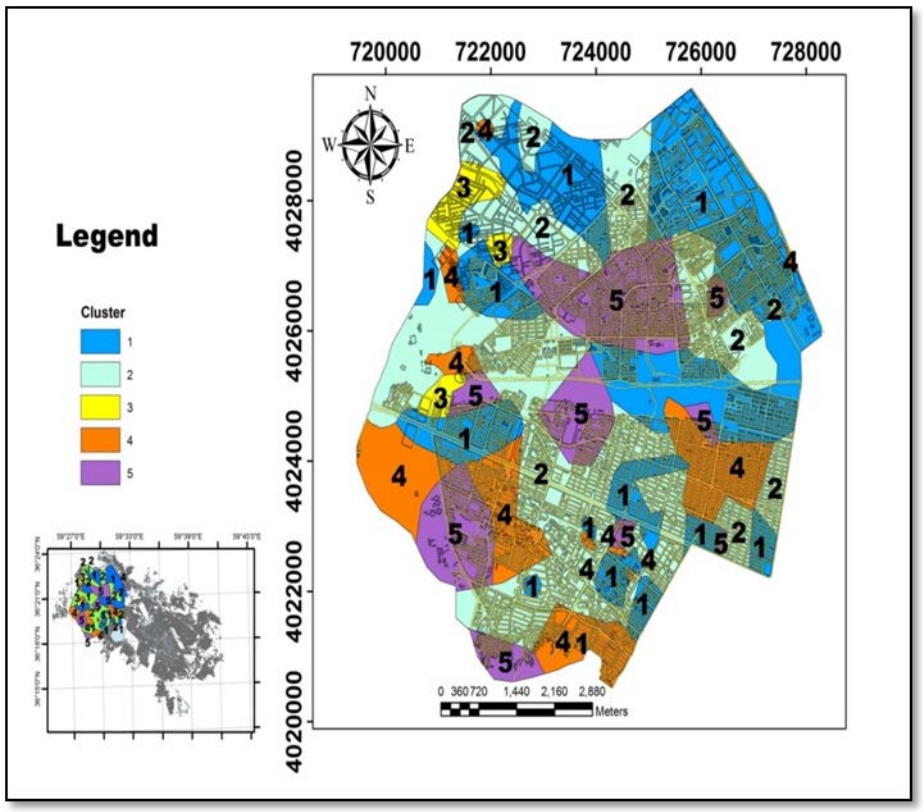


Figure 9. Soil texture map based on the clustering performed in this study (attention: no.5 is related to regions without data)

4.4. Multivariable regression

In multiple regression, the equation which predicts response variables as a linear function of p will be estimated by data. Statistically multi variable regression equations can be written as bellow in which Y is response variable and $X_1,\ldots, X_p$ are independent variables, equation 6.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e \qquad (6$$

The error term (e) is a vector of accidental variables and by assuming e being normal, it will be a function of mathematical expectation and variance of these accidental variables. And vector β is estimated such that sum of squared residuals is minimum.

Regression meaningfulness test

This test is for determination of existence or absence of a linear relation between Y and dependent variables of $X_1,\ldots,X_p$, equation7.

$$H_0: \beta_1 = \cdots = \beta_k = 0 \qquad (7$$

$$H_1: \beta_j \neq 0, j = 1,2,\ldots,k$$

Rejection of zero hypothesis means that at least one of the variables of $X_1,\ldots,X_p$ has a meaningful cooperation in the model and it infers the dependence of sum of squared residuals and sum of squared regression to each other. But if the zero hypothesis is correct it shows independency of sum of squared residuals and sum of squared regression from each other.

1- Selection of the best regression equation.

Two mutual criterions are involved in the selection of regression equations:

1- In order to make the equation appropriate for prediction purposes, we may want it to involve maximum possible X_s such that we can determine valid fitted values.

2- Because achieving data and preparing them for all X_s is costly we may want the equation to include minimum possible X_s.

3- Compromise between the two criterions is something that is generally called selection of the best regression equation and there are several methods for this operation including passro elimination method, step by step method and pishro method.

The most applicable method for constructing the variables is step by step method. In this method of entering every data into the model, all the variables formerly entered into the model which do not predict meaningfulness are

excluded from the model; it means that variables which their importance would decrease by the addition of other variables are deleted from the model. This method is a combination of pasro and pishro elimination method.

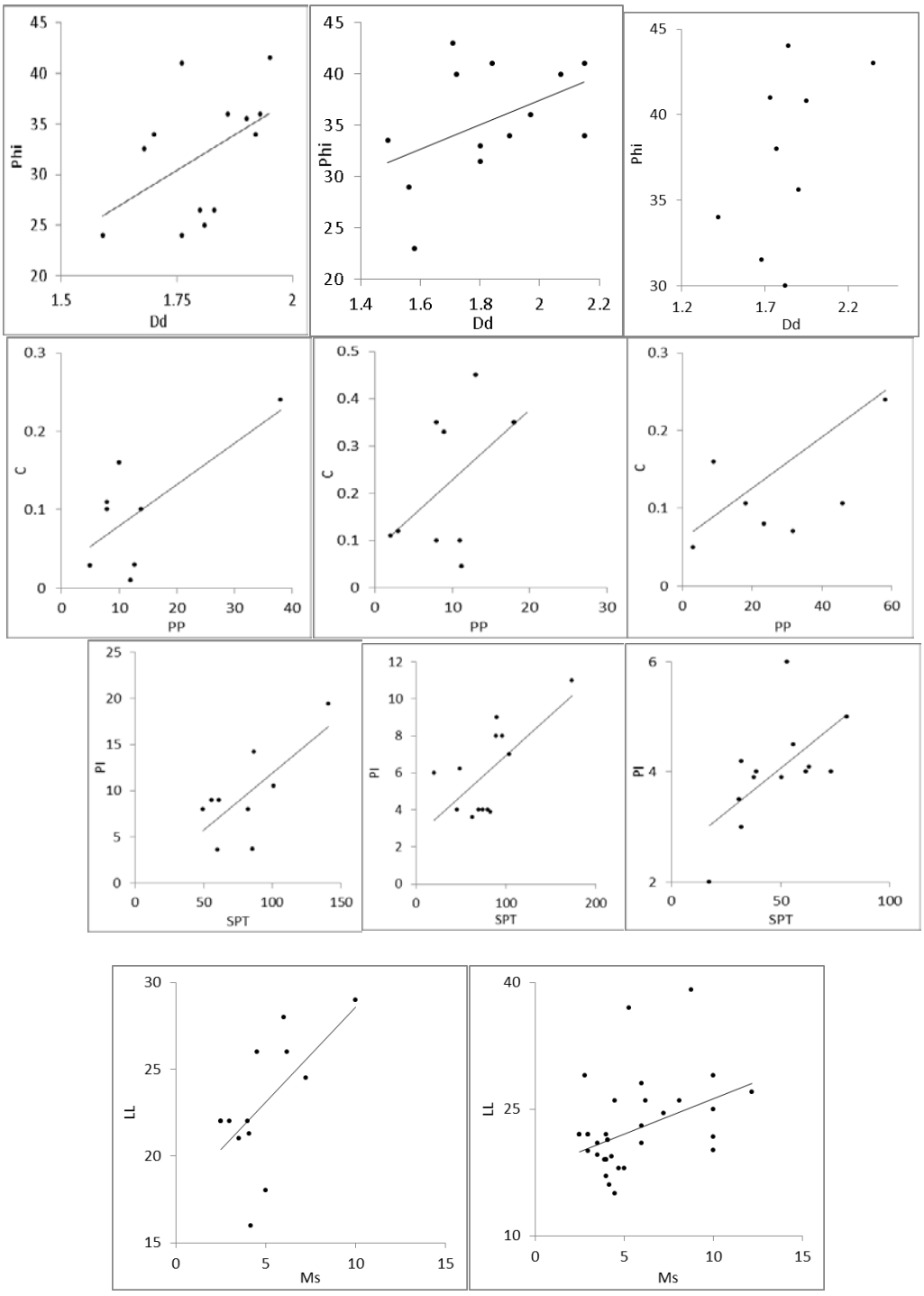


Figure 10. Linear regression between some of the most important physical parameters of western Mashhad soils with respect to different clusters.

Step by step method

In this method, two criterions are needed: one for entering variables and the other for variable deletion from the

model. primarily we enter the first variable by Forward Selection method (a variable which creates the most amount of change in R^2 and this variation in R^2 is to that extent that we can reject zero hypothesis which says the real value of variations equals zero,) and then by Backward Elimination selection method (a variable which creates minimum variations in R^2 value is deleted from the model and this variation in R^2 should be to that extent that do not reject the hypothesis which says the real value of R^2 variation equals zero) we analyzed both of them to see if none of them have the deletion criterion of the model and if it is not so then it will be deleted from the model. Hence in each step we enter a variable into the model based on the forward principles and then based on backward elimination rule, we analyze all the variables which are entered into the model heretofore; if they have the required criterion they will be deleted from the model. It should be noticed that meaningfulness level for entering a variable should be smaller than the meaningfulness level of excluding a variable; otherwise we would reach to a stage that a variable enters each time and is excluded again and this process continues to infinity if the computer is not stopped. In continue, results of this method are presented for test samples.

In equation 8, Dd is dry density, Phi is internal friction angle and PP is fine-grained percentage.

$$Dd = 0.13Phi - 0.004PP + 1.368 \quad (8$$

In this equation 9, C is cohesion, PP is fine-grained percentage, Phi is internal friction angle and Ms is saturation percentage.

$$C = 0.007PP - 0.006Phi - 0.030Ms. + 0.383 \quad (9$$

In equation10, PI is soil plasticity index and LL is Soil depth. Since the amount of liquid limit moisture in soils of this region is within 40 to 10 percent, the above equation clearly shows that the dominant clay soil type in this region of Mashhad is non-organic clay with low to average plasticity properties.

$$PI = 0.320LL - 1.120 \quad (10$$

Conclusion

The database used for this research includes 180 boreholes in western region of Mashhad. From the total of 180 boreholes, 3600 data sets were obtained with the introduced variables in the article. Results of the statistical analysis on this scattered population of data show that despite selecting sets of 1063 members with normal distribution is the only required condition for meaningfulness of the investigation results but this hypothesis is completely false for what is shown in this investigation. Because soil type can change the results as an effective variable, hence it's necessary to either perform a separate statistical analysis for every soil type or statistically analyze soils with similar physical performance altogether someway. The second way of the research which resulted in the creation of special characters for every cluster which are referred in the article, characterizes general results as below:

Existence of different soils in one group shows intensive variations in sedimentation of the region. Since this region is part of Kashafrood and Ghareh Ghoom basins, this matter doesn't seem illogical.

More than 80 percent of investigated samples in this research have less than 10 percent muck and clay which itself shows that the sedimentation area is full of energy. Small amount of clay and mock has caused this value to have very little impact on clustering.

A parameter which has more share in placing soils in similar groups is grain size and uniformity and this can mean similarity of sediment materials.

Since SPT values generally depend on grain size in other words values of parameters increase by increasing the grain size, therefore soils of this region of Mashhad are loose to semi-compacted. It is necessary to remind that soil stiffness can have an impact on SPT itself regardless of grain size, a status which cannot be observed very much in soils of this region.

Geotechnical database of western Mashhad is a relational database which is created by Access Software for data storage and retrieval. These data include physical, chemical and mechanical properties of soil which are obtained from different project reports. This database includes data of nearly 191 boreholes around the city.

Other benefits of the database include the ability to search different parameters especially in regions where soil data are not present.

Existing maps in the database are prepared in ArcView Software and are connected to the database. In order to preserve the robustness of data, a password was set for users to gain access to create, edit and delete data.

References

1. Maliki, M., 2002. Investigation of Geologic Properties, urban engineering and preparation of geotechnical maps of quaternary deposits of Tehran by preparing and developing Geotechnical data bank of Tehran (MSc thesis of geology engineering, Tarbiat Moalem university of Tehran.
2. Suwanwiwattana, P., Chantawarangul, K., Mairaing, W., Apaphant, P, 2001. The development of geotechnical database of Bangkok subsoil using GRASS-GIS, 22nd Asian Conference on Remote Sensing, Singapore.

3. Swift, JN., Stepp, JC, 2002. Development of archiving and web dissemination of geotechnical data, SMIP02 Seminar Proceedings, 161-175.
4. Vahaaho, I., Korpi, J., Hatva, E. M, 2003. Use of existing geotechnical information in urban planning, Land Use & Spatial Planning in Irland.
5. Yongli Gao, E., Calvin Alexander, J., Tipping, R.G, 2002. The development of a karst feature database for southeastern Minnesota, Journal of Cave and Karst Studies, 64 (1): 51-57.
6. Yusefi, A., 2003. Geological- geotechnical engineering investigation of city train of Mashhad, MSc thesis of geology.
7. Lugo Cintron, CY., 2007. Development of a geotechnical database for thecity of Mayaguez, Puerto Rico, Master's thesis in civil engineering, University of Puerto Rico.
8. Luna, R., Hertel, TP., Baker, H., Fennessey, T, 2001. Geotechnical database for emergency vehicle routes in Missouri, Proceeding of the 80th Annual Meeting of the Transportation Research Board, NRC, Washington D. C, CD ROM.
9. May, JH., 1999. Corcoran, M.K, Design and implementation of a comprehensive geotechnical database, US army engineer research and development center water-ways experiment station, Project summary.