Analyzing different factors responsible for closure of dug wells across various parts of India

1.Ranjan Gsk 2..Kailash Kolluru

*Bits Pilani, Hyderabad, India*

1.f20160245@hyderabad.bits-pilani.ac.in

2.f20160210@hyderabad.bits-pilani.ac.in

*Abstract*—This paper focuses on analyzing different factors responsible for closure of dug wells across various parts of India. The data has been taken from Open Government data platform . The purpose of this paper is to get key insights from the data set that might be useful for future schemes.

Keywords—python , pre-processing , correlation , grouping ,

linear encoding , matplotlib , standard deviation , co-variance

# Introduction

Dug wells are holes in the ground dug by shovel or backhoe. Historically, a dug well was excavated below the groundwater table until incoming water exceeded the digger's bailing rate. The well was then lined (cased) with stones, brick, tile, or other material to prevent collapse. Dug wells form the main source of water for many villages in India. Therefore. proper usage of it is important for the people for their well being.Wells have traditionally been sunk by hand digging, as is the case in rural areas of the developing world. These wells are inexpensive and low-tech as they use mostly manual labour, and the structure can be lined with brick or stone as the excavation proceeds. Drilled wells can access water at much greater depths than dug wells.

# Understanding data

The dataset has eighteen columns starting with states. Each state is again divided into its respective districts. Each district is divided into blocks which in turn are divided into villages. The remaining fourteen columns have the factors responsible for permenant stoppage in usage of dug wells and each row has the value which gives the number of such dug wells . Along with the number of dugwells the data also gives an insight into the potential loss of irrigation because of stoppage of these dugwells. This area is given in hectares.

The dataset is quite sparse which tells us that different villages have different factors for not using dug wells. The reasons for not using dug wells according to the given dataset are

* Salinity
* Dried up state
* Destroyed beyond repair
* Sea water intrusion
* Industrial effluents
* Availability of major/medium irrigation projects

# Pre-processing

Before we begin to work with our data we need to pre-process it to get accurate results.Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

The data is present in .xls format and we need to import it our python workspace. We do this with the help of python library called pandas . We import the data to our workspace as a dataframe. A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.

To start with pre-processing we first remove the columns that are not of importance to us when we analyze the data. In this case we drop district , block, village . However to find the correlation we form another data frame where we drop state, district ,block . For the initial dataframe we try to find all the unique states that are present in the dataset. We do this to group all the states together and to sum up their values for better analysis of results. Since state is a categorical variable we map it with an integer to use it in our analysis. Categorical variables cannot be used directly in machine learning algorithms. This process is also called Label encoding.The numpy library was used to achieve this purpose.

The label encoded values are grouped and their sum is considered . Apart from these columns two new columns are added for better visualization. One of the column is the total number of dugwells not being used in each state and the other one is total area in hectares which is being effected because of it.

Moving to the second dataframe , this dataframe is used to see if there is any correlation between different factors responsible for non usage of dug wells. For this purpose we don’t group the data according to states as viewing the data cumulatively for each state might give use misleading results as the water bodies vary from village to village.

The statistical relationship between two variables is referred to as their correlation. A correlation could be positive, meaning both variables move in the same direction, or

negative, meaning that when one variable’s value increases, the other variables’ values decrease. Correlation can also be neural or zero, meaning that the variables are unrelated.

* Positive Correlation: Both variables change in the same direction.
* Neutral Correlation: No relationship in the change of the variables
* Negative Correlation: Variables change in opposite directions.

To understand correlation we need to understand covariance. Variables can be related by a linear relationship. This is a relationship that is consistently additive across the two data samples.This relationship can be summarized between two variables, called the covariance.

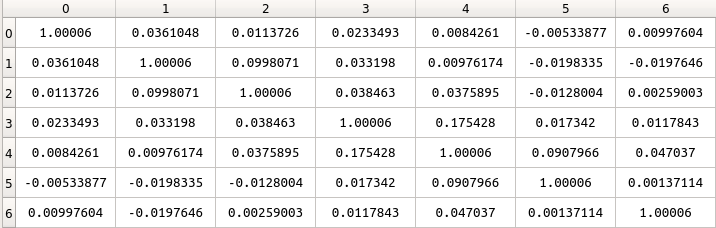
The calculation of the sample covariance is as follows:

**cov(X, Y) = (sum (x - mean(X)) \* (y - mean(Y)) ) \* 1/(n-1)**

To calculate the correlation using covariance we use the Pearson’s correlation coefficient which can confirm strength of linear relationship between two factors.

**Pearson's correlation coefficient = covariance(X, Y) / (stdv(X) \* stdv(Y))**

This constant is used to calculate the correlation among the various factors in our dataset . The values of correlation can be seen in Fig. 1.

1. Correlation Matrix

In the above correlation matrix numbers 0,1,2....,6 refer to the factors responsible for stoppage dug well in the same order. From the above matrix we can infer that none of the factors are correlated strongly as almost all of the values are very close to zero.

Hence all the factors will be used for further analyzing the data. Here the irrigation potential lost has not been used as the number of dugwells lost due to a particular factor will be directly proportional to the amount of area lost . Hence calculating correlation for irrigation potential lost will give us the same result approximately.

# Visualization

The dataframe which was grouped based on states and summed was used for the visualization purpose . Matplotlib was used to show the plots. Total eight plots were plotted for the purpose of data analysis. Each plot has number of dug wells permenantly stopped on x-axis and the area of potential irrigation lost on y-axis. For every factor a graph has been plotted with the scatter points as the name of the states.

These plots help us to understand which state has been the most affected due to a a particular factor and at the same time which state has been the least affected one. An additional graph has been plotted which uses the total number of dugwells not functioning and the total area of potential irrigation loss. Figure 2 gives an example of one such plot

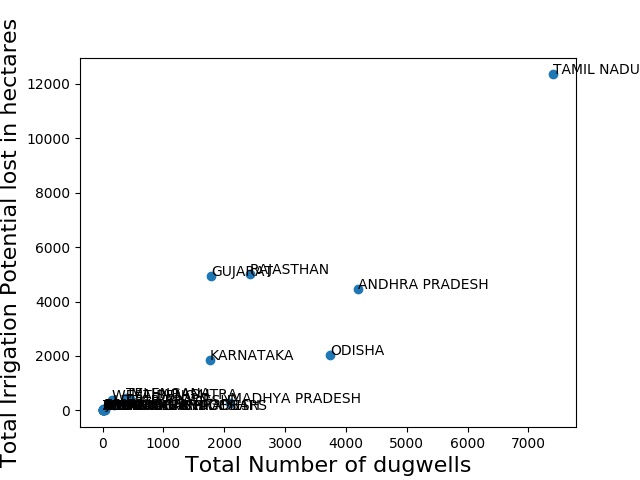


Fig.2.Dugwell Plot

From the above plot we can clearly understand that the overall affect of all the factors has been highest on Tamil Nadu followed Gujarat,Rajasthan and Andhra Pradesh.

# Clustering

Clustering is one of the most widely recognized exploratory information examination procedure used to get an instinct about the structure of the information. It very well may be characterized as the errand of distinguishing subgroups in the information with the end goal that information focuses in a similar subgroup (bunch) are fundamentally the same as while information focuses in various groups are altogether different. As such, we attempt to discover homogeneous subgroups inside the information with the end goal that information focuses in each bunch are as comparable as conceivable as indicated by a similitude measure, for example, euclidean-based separation or relationship based separation. The choice of which closeness measure to utilize is application-explicit.

Clustering investigation should be possible based on highlights where we attempt to discover subgroups of tests dependent on highlights or based on tests where we attempt to discover subgroups of highlights dependent on tests. Clustering is utilized in showcase division; where we attempt to fined clients that are like each other whether as far as practices or traits, picture division/pressure; where we attempt to assemble comparable districts, report grouping dependent on themes, and so on.

In contrast to regulated picking up, clustering is viewed as a solo learning technique since we don't have the ground truth to analyze the yield of the bunching calculation to the genuine marks to assess its exhibition.

# K-MEANS

K means algorithm is an iterative calculation that attempts to parcel the dataset into K pre-characterized unmistakable non-covering subgroups where every datum point has a place with just one gathering. It attempts to make the between group information focuses as comparable as could reasonably be expected while additionally keeping the clusters as various as could be expected under the circumstances. It relegates information focuses to a bunch with the end goal that the total of the squared separation between the information. Its an Prototype based clustering where each cluster is represented by a prototype called Centroid(average of similar points for continuous features).Each Point is assigned to a cluster with closest centroid. This clustering needs an predefined no of clusters K.

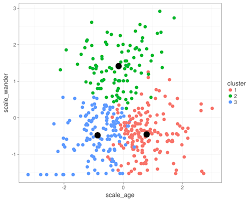


Fig.3. K Means example

The way k means calculation works is as per the following:

* Determine the value k that is the number of clusters.
* Introduce centroids by first rearranging the dataset and afterward arbitrarily choosing K information focuses for the centroids without substitution.
* Continue emphasizing until there is no change to the centroids. i.e task of information focuses to bunches isn't evolving.
* Figure the entirety of the squared separation between information focuses and all centroids.
* Allocate every point to the nearest centroid.
* Process the centroids for the clusters by taking the mean of the all information focuses that have a place with each group.

Since clustering algorithms including kmeans use separation based estimations to decide the comparability between information focuses, it's prescribed to normalize the information to have a mean of zero and a standard deviation of one since quite often the highlights in any dataset would have various units of estimations, for example, age versus pay. The min max normalization technique has been used for the dataset.

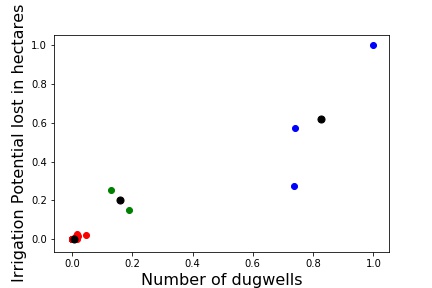
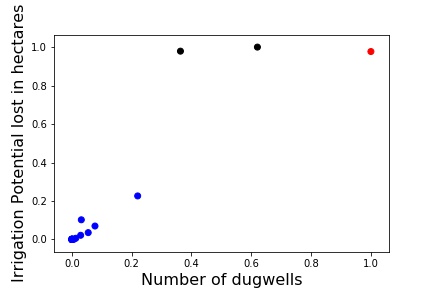


Fig.4. K Means on one of the factor

Fig.4 gives the plot of K means on one of the factor of the dataset. The black circles indicate the centroids. The three clusters are colored as blue green and red. The cluster number has been chosen by using the WCSS technique.



# DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning. DBSCAN groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

**Parameters:**

The DBSCAN algorithm basically requires 2 parameters:

**eps**: specifies how close points should be to each other to be considered a part of a cluster. It means that if the distance between two points is lower or equal to this value (eps), these points are considered neighbors.

**minPoints**: the minimum number of points to form a dense region. For example, if we set the minPoints parameter as 5, then we need at least 5 points to form a dense region*.*

**Parameter estimation:**

**eps**: if the eps value chosen is too small, a large part of the data will not be clustered. It will be considered outliers because don’t satisfy the number of points to create a dense region. On the other hand, if the value that was chosen is too high, clusters will merge and the majority of objects will be in the same cluster. The eps should be chosen based on the distance of the dataset (we can use a k-distance graph to find it), but in general small eps values are preferable.

**minPoints**: As a general rule, a minimum minPoints can be derived from a number of dimensions (D) in the data set, as minPoints ≥ D + 1. Larger values are usually better for data sets with noise and will form more significant clusters. The minimum value for the minPoints must be 3, but the larger the data set, the larger the minPoints value that should be chosen.

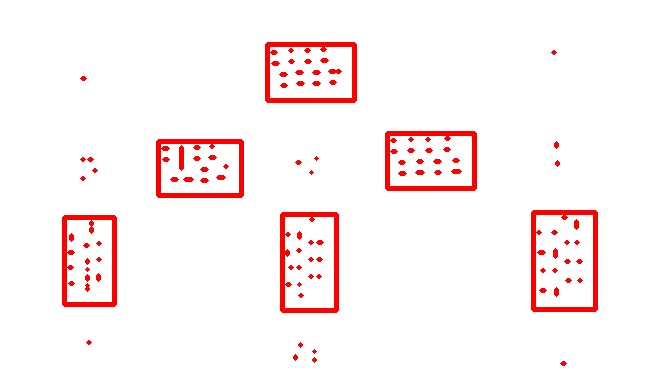
The parameter estimation is a problem for every data mining task. To choose good parameters we need to understand how they are used and have at least a basic previous knowledge about the data set that will be used.

Fig 6. Plot of dbscan on one of the factors

Fig 6. Shows the plot of dbscan on one of the factors. The different cluster points are marked with different colours. The data which does not belong to any of the cluster is considered as noise in DBSCAN. The eps value is chosen as 0.3 and the minPoints value is set to 2 to achieve proper results.

1. CONCLUSION

The main conclusion from the above clustering will be to group together the states that are the most, least and moderately affected by the closure of the dugwells. The plots give us visualization of the states are scattered with number of dugwells on x-axis and irrigation potential lost in hectares in y-axis.An additional graph has been plotted which uses the total number of dugwells not functioning and the total area of irrigation potential lost.

##### References

1. *https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/*
2. [*https://en.wikipedia.org/wiki/Well*](https://en.wikipedia.org/wiki/Well)
3. [*https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80*](https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80)

Fig 5. Example of DBSCAN