International Journal of Innovation Engineering and Science Research

Editorial Board

International Journal of Innovation Engineering and Science Research

Volume 7 ISSUE 1 January-February 2023

International Journal of Innovation Engineering and Science Research
(ISSN: 2581-4591)

www.ijiesr.com

Parallel Feature Selection approach for Big Data Processing using Coarse-grained Genetic algorithm

Waad Bouaguel
College of Business, University of Jeddah

ABSTRACT

The Genetic Algorithm (GA) is one of the most popular meta-heuristic algorithms that has been found to be effective for ac- curate feature selection tasks. In recent decades, the number of instances and features hasgrown in size, making most current feature selection algorithms and even GAs, unscalable. To improve the scalability of big data feature selection algorithms and exceptionally GA, we adopt distributed computing strategies such as the Map-Reduce model and Hadoop system. In this paper, we first present recent work dealing with the use of parallel genetic algorithms in large datasets. Then we will propose a new parallel genetic algorithm based on the island model. The performance of the proposed method was theoretically and empirically compared to existing feature selection methods when handling large-scale datasets and the results confirm the effectiveness of our pro- posed method.

Keywords—*Genetic algorithm; featureselection; island model; parallel processing.*

I. INTRODUCTION

Feature selection is a category of dimensionality reduction. In this category, a subset of existing features is selected without a transformation [1]. The idea behind feature selection is to reduce the influence of tricky features in a dataset, where tricky and unnecessary features include irrelevant and redundant features. Removing these features reduces the dimensionality of the search space and speeds up the learning algorithm[2], [3].

The selection of the optimal feature subset is an optimization problem that proved to be NP-hard. Two major approaches are traditionally used to deal with NP-hard problems, exact methods, and meta-heuristics [4]. Exact methods allow the exact solution to be found, but this approach is impractical since it is extremely time-consuming in real-world problems. On the other hand, meta-heuristics are used for solving complex and real-world problems, because they provide sub-optimal solutions in a reasonable time.

GA is an adaptive heuristic search algorithm inspired by the process of natural evolution and one of the most well-known Population-based meta-heuristics. As reported by [5], GA was faster in finding near-optimal features from large datasets compared to other algorithms. Despite the efficiency of GA in feature selection, as most of the existing feature selection algorithms, GAs is also proposed to deal with small data sets under several gigabytes and become impractical in Big data cases [6], [7]. Hence, to overcome this kind of issue feature selection is performed in a distributed manner. This is a simple task for GAs since GA is naturally parallelizable since it searches simultaneously from multiple points, not from a single point. Three levels of parallelization can be exploited: Global parallelization model, Fine-grained parallelization model, and Coarse-grained parallelization or island model [8]–[10]

The global parallelization model uses one master node to apply genetic and selection operators and distributes individuals among the remaining slaves, where the fitness values of the individuals are computed [11]. This model provides lower run time only for very slow objective functions which is a drawback according to [12] and its search mechanism uses a single population. The fine-grain parallel GA uses an overlapped neighborhood system to provide a smooth diffusion of the individuals [13], [14], while the coarse-grained model also known as the island model distributes the population among the processors. Eachnode called an island is given a sub-population to process [13]. The genetic Algorithm is then applied toeach sub-population. The individuals are then migrated to other nodes according to some given criteria in order to synchronize the solution set. This model has two main advantages:

(1) sub-populations could explore various portions of the search space and (2) there is less communication between nodes [15].

In this paper, we present a new Parallel GA based on the island model for feature selection. The performance of the proposed method will be theoretically and empirically compared to existing feature selection methods. This paper is organized as follows: in section 2, we discuss recent works on Parallel GA for large datasets and feature selection. Section 3 describes the proposed method. Section 4 is dedicated to presenting the experimental study and Section 5 gives the results and discussion.

II. RECENT WORKS ON PARALLEL GA FOR LARGE DATASETS AND FEATURE SELECTION.

In recent years, evolutionary algorithms have gained popularity in solving complex problems. Several research works have been done in this context, including a study by [16] that used a weighted Gene Genetic Algorithm for big data feature selection. The proposed method stores a weight for each gene on the population solutions. Then, the stored weights are decreased or increased during the evolution of the genetic algorithm based on the appearance of the tested set of features on the best solutions. The proposed algorithms were applied to five datasets and the results show that the proposed algorithms can effectively enhance the classifying performance against the other feature selection algorithms. In another work, [17] performed feature selection through the implementation of a selection process by a genetic algorithm, using different classification techniques, which allows the generation of a classification model of a set of children's activities based on the environmental sound.

[8] defines a framework for implementing parallel Genetic Algorithms (GAs) on the Hadoop platform by applying the MapReduce model.[18] proposed a parallel GA to derive bi-clusters from the microarray matrix, and [19] proposed a wrapper technique using KNN classifier for supervised feature selection to yield good classification accuracy with minimal features. To improve the performance, they used a master-slave Parallel Genetic algorithm using Hadoop MapReduce and they suggested a wrapper technique using K-NN classifier for supervised feature selection to yield good classification accuracy with minimal features. [20] proposed a MapReduce implementation of a parallel GA to find the minimal rough set redact. The model targeted many components such as constructing a distinction table, GA population evaluation, and GA operations. Several mappers have been used to construct the distinction table and the reduce phase is charged with the genetic operations. The driver is charged with collecting the results to select the best individual. This approach has reduced the execution time without demeaning the solution quality in terms of the redact size. Using parallel also programming, a Fuzzy pre-processing technique was used to reduce the input data for a Multi-objective GA. A Parallel SVM Classifier was used to assess the best-identified gene subsets [21].

III. PROPOSED METHOD

Step I: In order to perform GA, an initial population of features is created by randomly selecting a set of ordered rank lists of features. From empirical studies, over a wide range of problems, a population size between 30 and 100 is usually recommended [22].

Step II: Once the initial population is fixed, we need to select new members for the next generation. To do this, we can use the MapReduce model, which divides the whole population into several separate sub-populations. the whole population is divided into smaller blocks, and then the HDFS files are used to create new populations [23].

Step III: This step is called the map phase. The mapper function processes the input data and creates several small chunks of data.

GA is applied to each data chunk and the resulting map output for each dataset of m features is m pair of key-value < k; v > that refers to the < features; values of features >. In this phase, we apply the genetic operations selection, crossover, and mutation, to each sub-population, which means different sub-populations will take different crossover rates and mutation rates. After every 100 generations, the best individual in each sub-population will be sent to other sub-populations and replace the worst individual. The selected members are then crossed over with the cross-over probability. Therefore, crossover combines the features of two lists to create two similar ranked lists. The mutation operator is used, it acts as a population perturbation operator. Mutation operates by randomly changing one or more elements of any list. Typically, the mutation does not occur frequently, so the mutation is of the order of 0.001 [22].

Step IV: The map phase output contains the features selected by the GA, and the reduce phase aggregates all the relevant features in one file.

IV. EXPERIMENTAL STUDY

A. Dataset description and experimental environment

We used four datasets from the UCI Machine Learning Repository to assess the performance and effectiveness of our proposed method. These datasets were the Madelon dataset, the Semeion Handwritten Digit dataset, the Hill-Valley dataset, and the Breast Cancer dataset.Our approach was implemented using the Hadoop MapReduce implementation with R language on Amazon Web Services Cloud, and the RHadoop project that brings several R packages to work with Hadoop interactively.In this work, we deal with three kinds of implementation: a traditional GA with no parallel processing, a Parallel GA on Hadoop platform, and a fully distributed version on the Amazon cluster.

TABLE 1 Clearly illustrates the experimental environment for each implementation.

- Nonparallel GA: GA was used for the feature selection task with no parallel processing on a single machine with 6 GB of RAM.
- Parallel GA on Hadoop platform: the solutions are represented in the form of a population shared globally and managed by a master processor. The other slave processors take the charge of evaluating the fitness of the global population. the proposed approach was implemented over a Hadoop platform on a single machine with 6 GB of RAM.
- Parallel GA with an Amazon cluster: required a full Hadoop cluster composed of 1 master and 3 slaves, each machine has 2 CPUs and 8 GB of RAM.

TABLE I. EXPERIMENTALENVIRONMENT.

	Non-Parallel GA	***Parallel GA* Hadoop platform**	***Parallel GA* *Amazon cluster***
Number of machines	One machine	One machine	One master and three slaves
RAM	6 GB	6 GB	8 GB
Framework		Hadoop 2.7.3	Hadoop 2.7.3
Language	R (3.3.1)	R (3.3.1)	R (3.3.1)

B. Parameters setting

In this section, we aim to understand the impact of the GA's parameters on the learning results, and how they affect the trade-off between fast convergence and exploratory power.

1) Setting for Crossover operator and rate

In GAs, the crossover operation is used to create new individuals from the old ones, thus making them more likely to be better than the parents. Some well-known crossover operators are reflected in [24] and summarized in TABLE 2

TABLE II. SUMMARY OF CROSSOVER OPERATORS

Singlepoint; n-point; Discrete; Arithmetical; Simulatedbinary; Unimodalnormaldistribution; Parentcentric; Wright'sheuristic; Dynamicfuzzy; connectivebased simplex; Averagebound; Geometrical; Uniform

We propose to use different crossover operators in each generation of the GA to provide diversity in the population and improve the overall performance of the GA. Hence in each generation, a different crossover operator is chosen from the set of operators presented in TABLE 2

2) Setting of the Mutation operator

In this part we try to apply multi-mutation operators to a GA, so that in each generation a different mutation operator is chosen, thereby increasing diversification and enhancing the GA's performance. Several mutation operators were discussed in [24] and presented in TABLE 3.

3) setting of Crossover and mutation probability rate

Crossover occurs with a probability pc and mutation occurs with a probability pm. The crossover rate and mutation rate control the behavior and performance of GAs.

The higher the crossover rate, the quicker exploitation proceeds. A pc that is too large would disrupt individuals faster than they could be exploited. Small pm values are commonly adopted in GAs. Typicalvalues of pc are in the range 0.5 -1.0, while typical values of pm are in the range 0.001-0.05.

TABLE III. SUMMARY OF MUTATION OPERATORS

Single point; n-point; Discrete; Arithmetical; Simulated binary; Unimodal normal distribution; Parent centric; Wright's heuristic; Dynamic fuzzy; connective based simplex; Average bound; Geometrical; Uniform

To identify suitable values of GA parameters, the performance of the GA will be evaluated for several different sets of these parameters, see Table 4.

TABLE IV. GA CONTROL PARAMETER SETS. CONTROL

Parameter	***Values***
Number of generations	[50;100]
Population size	[10;20;30]

Parameter	***Values***
Probability of Crossover	[0.8;1.0]
Probability of Mutation	[0.01;0.02]

To identify suitable values of GA parameters, 24 different sets of GA control parameters will be tested. The accuracy of the classification results will be improved by using severalgenerations of 100, a population size of 20, a crossover rate of 0.8, and a mutation rate of 0.1, these values will be used in the rest of the experiments. see Table 5 for more details.Each parameter set shown in Table 5 was run 10 times, and each time

with different initial starting search points to eliminate the effect of initial randomness in the selected solutions.

TABLE V. TABLE5.ACCURACYSUMMERFOREACHSEATOFPARAMETERS

(NUMBEROFGENERATION,POPULATIONSIZE,CROSSOVERRATE,MUTATIONRATE)

Index	Parameter set	Accuracy	Index	Parameter set	Accuracy	Index	Parameter set	Accuracy
1	[50;10;0.8;0.01]	0.571	2	[50;10;0.8;0.02]	0.578	3	[50;10;1;0.01]	0.582
4	[50;10;1;0.02]	0.567	5	[50;20;0.8;0.01]	0.560	6	[50;20;0.8;0.02]	0.601
7	[50;20;1;0.01]	0.600	8	[50;20;1;0.02]	0.570	9	[50;30;0.8;0.01]	0.459
10	[50;30;0.8;0.02]	0.462	11	[50;30;1;0.01]	0.502	12	[50;30;1;0.02]	0.494
13	[100;10;0.8;0.01]	0.490	14	[100;10;0.8;0.02]	0.632	15	[100;10;1;0.01]	0.590
16	[100;10;1;0.02]	0.487	**17**	**[100;20;0.8;0.01]**	**0.689**	18	[100;20;0.8;0.02]	0.588
19	[100;20;1;0.01]	0.561	20	[100;20;1;0.02]	0.572	21	[100;30;0.8;0.01]	0.602
22	[100;30;0.8;0.02]	0.569	23	[100;30;1;0.01]	0.604	24	[100;30;1;0.02]	0.490

4) Fitness function

The fitness function is one of the most important parameters of GA because it evaluates how good a solution is. In our case, the fitness function was chosen to be RandomForest (RF) [25], and the accuracy of RF helped to decide if a candidate feature subset fit or not. Hence, many sets of randomly selected features are generated from each file. Where, each one of them is represented by an F-bit string in which the value '1' or '0' of any bit means the presence or absence of the corresponding feature, respectively. For each generation, individuals are used to producing a forest of decision trees. Then a fitness score is assigned to each individual based on how well the corresponding tree classifier classified the test dataset using the root-mean-square.

V. RESULTS AND DISCUSSION

A. Evaluation of running time

The given results in Table (6) show that the proposed approach takes less time to choose the most relevant features compared to the non-parallel version and the parallel Hadoop implementation on a single machine.

TABLE VI. RUNNINGTIMERESULTS.

	Runningtime(s)	Selectedfeatures
	Madelon data set	
NonParallel GA	207945.992	261
ParallelGAonHadoopplatform	204530.196	238
ParallelGAwithanAmazoncluster	151949.403	246
	Semeion Handwritten Digit data set	
NonParallel GA	28490.67	81
ParallelGAonHadoopplatform	26954.098	110
ParallelGA with an Amazone cluster	16428.100	114
	Hill-Valley data se	
NonParallel GA	4839.744	37
ParallelGAonHadoopplatform	4092.932	40
ParallelGA with an Amazone cluster	2277.31	44

Table 6 show that for the Madelon dataset our approach success to select features in 151949.403 seconds compared to 207945.992 and 204530.196 second given respectively by the nonparallel and the single machine implementation. We also notice from Table 6 that the new approach with the Amazon cluster obtained also the smallest running time for the Semeion Handwritten Digit dataset. The same results were confirmed with the Hill-Valley dataset where the best set of features was selected in half the time used by the non-parallel GA and the single-machine implementation.

This improvement in running time for both parallel implementations confirm the importance of introducing parallel computing to GA paradigm. The proposed approach for feature selection in GA takes less time compared to the non-parallel version and the parallel Hadoop implementation on a single machine. The best running time was obtained with the Amazon cluster.

B. Evaluation of learning performance

In this part, we evaluate the proposed approach according to the learn- ing results. Accuracy and F-measure are used as classification measures.

From Table (7) and Table (8) we evaluate the effectiveness of our proposed feature selection method using SVM and RF classifiers. The obtained results show the effectiveness of selecting features using the island and the MapReduce models in improving the classification results.

Let's start with the results given by the SVM classifier for the Madelon dataset as shown in Table 7. The accuracy and F-measure of the SVM classifier for the Madelon dataset were 0.4855 and 0.58, respectively, for the non-parallel implement of GA and the parallel implementation on the Haddop platform. While they were about 0.533 and 0.59 for our proposed method for 246 selected features. The obtained results using our approach are higher than those provided by the non-parallel implement of GA and the parallel implementation on Hadoop platform.

From Table 7 we notice that the proposed approach achieves the best rates of classification for the Semeion Handwritten Digit dataset when compared to the baseline model. As shown in Table 7 we achieve the best rates of accuracy (.i.e. 0.992) and F-measure (.i.e. 0.990) and the smallest time for model building. For the Hill-Valley dataset, feature selection improved the classification performance

(accuracy and F-measure) and reduced the model-building time. This finding comes just confirms the importance of feature selection and its effect on improving classification performance.

TABLE VII. CLASSIFICATION RESULTS FOR USING SVM.

	Time	Accuracy	F-measure
		Madelon data set	
Allfeatures	2.440	0.485	0.58
Non-parallel GA(261 features)	0.408	0.488	0.57
ParallelGAonHadoopplatform (238features)	0.372	0.520	0.581
ParallelGAwithAmazoncluster (246features)	0.361	0.533	0.59
	Semeion Handwritten Digit data set		
Allfeatures	0.592	0.979	0.98
Non-ParallelGA(81features)	0.608	0.979	0.98
ParallelGAonHadoopplatform (110features)	0.348	0.991	0.99
Parallel GA withan Amazon cluster(114features)	0.398	0.992	0.99
	Hill-Valleydataset		
Allfeatures	0.340	0.502	0.260
Non-ParallelGA(37features)	0.048	0.508	0.262
ParallelGAonHadoopplatform (40features)	0.072	0.530	0.27
Parallel GA withan Amazon clus-ter(44features)	0.068	0.532	0.27

Time: Time for building the model in seconds

Let's move to the results given by the RF classifier and summarized in Table 8. The obtained results confirm our assumptions about the importance of introducing parallel computing in feature selection. From the obtained results we conclude on the effectiveness of the proposed approach that obtained not only the best learning performance but alsothe best reduced time for building the model.

TABLE VIII. CLASSIFICATION RESULTS USING RF.

	Time	Accuracy	F-measure
	Madelondataset		
Allfeatures	6.608	0.511	0.56
NonParallelGA(261features)	3.386	0.483	0.59
ParallelGAonHadoopplatform (238features)	3.088	0.515	0.61
Parallel GA withan Amazon cluster(246features)	3.262	0.526	0.61
	Semeion HandwrittenDigitdataset		
Allfeatures	716.608	0.712	0.88
NonparallelGA(81features)	226.004	0.895	0.874
ParallelGAonHadoopplatform (110features)	3.630	0.902	0.98
Parallel GA withan Amazon cluster(114features)	3.570	0.974	0.98
	Hill-Valleydataset		
Allfeatures	183.132	0.5138	0.48
NonParallelGA(37features)	55.64	0.5635	0.58
ParallelGAonHadoopplatform (40features)	0.860	1	1
Parallel GA withan Amazon cluster(44features)	0.872	1	1

Time: Time for building the model in seconds

VI. CONCLUSION

This paper investigates a new feature selection approach based on Genetic Algorithm using parallel processing. It evaluates the efficiency ofthe proposed method on real data sets from the UCI Machine Learning Repository. The obtained results confirm our assumptions about the importance of introducing parallel computing in feature selection. In the future, we intend to study in deep other parallel implementations of a large range of evolutionary feature selection algorithms.

REFERENCES

[1] F. Sahin, “A survey on feature selection methods,” Computers and Electrical Engineering, vol. 40, pp. 16–28, 01 2014.

[2] N. AlNuaimi, M. M. Masud, M. A. Serhani, and N. Zaki, “Streaming feature selection algorithms for big data: A survey,” Applied Computing and Informatics, 2019.

[3] I. Rodriguez, R. Huerta, C. Elkan, and C. S. Cruz, “Quadratic programming feature selection,” Journal of Machine Learning Research. 11, no. 4, pp. 1491–1516, 2010.

[4] W. Bouaguel, G. Mufti, and M. Limam, A New Feature Selection Technique Applied to Credit Scoring Data Using a Rank Aggregation Approach Based on: Optimization, Genetic Algorithm and Similarity, 03 2015, pp. 337–334.

[5] A. Alzubaidi, G. Cosma, D. Brown, and G. Pockley, “Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information,” 10 2016.

[6] Y. Li, T. Li, and H. Liu, “Recent advances in feature selection and its applications,” Knowledge and Information Systems, vol. 53, pp. 551–577, 2017.

[7] M. Rong, D. Gong, and X. Gao, “Feature selection and its use in big data: Challenges, methods, and trends,” IEEE Access, vol. PP, pp. 1–1, 01 2019.

[8] F. Ferrucci, P. Salza, T. Kechadi, and F. Sarro, “A parallel genetic algorithms framework based on hadoopmapreduce,” 04 2015, pp. 1664–1667.

[9] A. Natarajan and B. Ramasamy, “A fuzzy parallel island model multi objective genetic algorithm gene feature selection for mi- croarray classification,” vol. 11, pp. 2761–2770, 03 2016.

[10] Shyh-Chang Lin, W. F. Punch, and E. D. Goodman, “Coarse-grain parallel genetic algorithms: categorization and new approach,” in Proceedings of 1994 6th IEEE Symposium on Parallel and Distributed Processing, 1994, pp. 28–37.

[11] F. De Toro, J. Ortega, J. Fern´andez, and A. D´ıaz, “Psfga: a parallel genetic algorithm for multiobjective optimization,” in Parallel, Distributed and Network-based Processing, 2002. Proceedings. 10th Euromicro Workshop on. IEEE, 2002, pp. 384–391.

[12] E. Alba and J. Troya, “Troya, j.m.: Improving flexibility and effi- ciancy by adding parallelism to genetic algorithms. statistics and computing 12(2), 91-114,” Statistics and Computing, vol. 12, pp. 91–114, 04 2002.

[13] N. H., S.Mahajan, and S. Omkar., “Big data clustering using ge- netic algorithm on hadoop mapreduce,” International Journal Of Scientific and Technology Research, vol. 4, April 2015.

[14] L. Di Geronimo, F. Ferrucci, A. Murolo, and F. Sarro, “A parallel genetic algorithm based on hadoop mapreduce for the automatic generation of junit test suites,” in Software Testing, Verification

[15] and Validation (ICST), 2012 IEEE Fifth International Conference on. IEEE, 2012, pp. 785–793.

[16] J. Yazidi, W. Bouaguel, and N. Essoussi, “A parallel imple- mentation of relief algorithm using mapreduce paradigm,” in In- ternational Conference on Computational Collective Intelligence. Springer, 2016, pp. 418–425.

[17] T. Abed Mohammed, O. Bayat, O. Ucan, and S. Alhyali, “Hybrid efficient genetic algorithm for big data feature selection problems,” Foundations of Science, 03 2019.

[18] A. Garcia-Dominguez, C. E. Galvan-Tejada, L. A. Zanella- Calzada, H. Gamboa-Rosales, J. I. Galvan-Tejada, and J. M. Celaya, “Feature selection using genetic algorithms for the gen- eration of a recognition and classification of children activities model using environmental sound,” Mobile information systems, vol. 2020, 2 2020.

[19] A. Laishram and S. Vipsita, “Bi-clustering of gene expression mi- croarray using coarse grained parallel genetic algorithm (cgpga) with migration,” in India Conference (INDICON), 2015 Annual IEEE. IEEE, 2015, pp. 1–6.

[20] G. T. Hilda and R. Rajalaxmi, “Effective feature selection for supervised learning using genetic algorithm,” in Electronics and Communication Systems (ICECS), 2015 2nd International Con- ference on. IEEE, 2015, pp. 909–914.

[21] M. A. Alshammari and E.-S. M. El-Alfy, “Mapreduce implemen- tation for minimum reduct using parallel genetic algorithm,” in Information and Communication Systems (ICICS), 2015 6th In- ternational Conference on. IEEE, 2015, pp. 13–18.

[22] A. Natarajan and R. Balasubramanian, “A fuzzy parallel island model multi objective genetic algorithm gene feature selection for microarray classification.” International Journal of Applied Engineering Research, vol. 11, no. 4, pp. 2761–2770, 2016.

[23] V. Pihur, S. Datta, and S. Datta, “RankAggreg, an R package for weighted rank aggregation,” BMC Bioinformatics, vol. 10, no. 1, pp. 62–72, 2009.

[24] Q. He, X. Cheng, F. Zhuang, and Z. Shi, “Parallel feature selection using positive approximation based on mapreduce,” in Fuzzy Sys- tems and Knowledge Discovery (FSKD), 2014 11th International Conference on. IEEE, 2014, pp. 397–402.

[25] S. L. Siew Mooi, A. B. Md Sultan, M. Sulaiman, A. Mustapha, and K. Y. Leong, “Crossover and mutation operators of genetic algorithms,” International Journal of Machine Learning and Com- puting, vol. 7, pp. 9–12, 02 2017.

[26] G. Biau, F. C´erou, and A. Guyader, “On the rate of convergence of the bagged nearest neighbor estimate,” Journal of Machine Learn- ing Research, vol. 11, no. Feb, pp. 687–712, 2010.

[27] .

International Journal of Innovation Engineering and Science Research
(ISSN: 2581-4591)
www.ijiesr.com

Soft Start of Three-Phase Transformers

[1]**Mohamed A. Shaban**
College of scinecs and Technology
Power Engineering
Surman, Libya

[2]**Salah Eddin Albakoush**
College of scinecs and Technology
Power Engineering
Surman, Libya.

[3]**Adel Ali Gamati**
College of scinencs and Technology
power Engineering
Surman .Libya

ABSTRACT

Soft-starters of three-phase transformer. This article deal with nowadays used methods for three-phase transformer inrush current suppression. Second part of the article discussed possibility of application non-informed methods (no the remnant flux measuring) of one-phase transformers inrush current suppression to three phase transformers. Measurement results of the considered controlled switching method are published at the end of article. There was tested three-phase transformer Legrand 10 kVA.

Keywords—*Three pahse, transformer, control switching.*

I. INTRODUCTION

The inrush current of three-phase transformers arises as a result of their oversaturation magnetic circuit similarly to single-phase transformers. Oversaturation occurs when at the moment the transformer is connected to the power supply; the remnant magnetic flux in the flux core does not correspond in steady state. A more detailed explanation is provided in a number of sources, e.g. in [1].In the case of three-phase transformers, the analysis of the origin and solution of the inrush current is complicated by the great variability of the structural arrangements of the magnetic circuit common to all phases and by the number of group winding connections. At the same time, the various variants differ considerably. In this article they are methods presented for the connection of the primary windings into a star with a drawn center. Some methods however, they are also applicable to other group connections.

II. THREE-PHASE SWITCHING CURRENT LIMITATION METHODS TRANSFORMERS

Three-phase transformers are mostly used for larger outputs, many of the principles of limiting the inrush current for single-phase transformers, which are suitable for smaller performances,

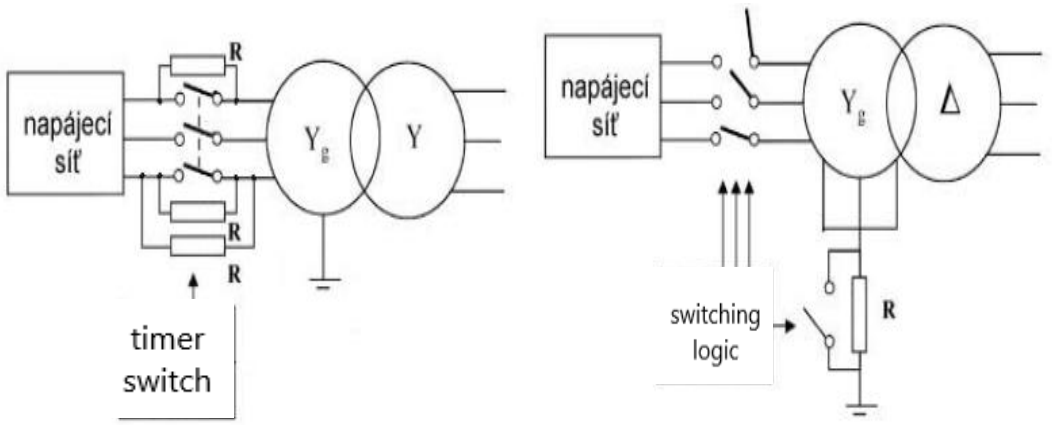

Fig .1: On the left, a resistive starter, on the right, a resistive starter in the middle conductor in combination with sequential switching according to [4, 5]

Several of the following methods appear in practical implementations and professional Publications. The standard method is to start with a resistance or induction starter. The transformer is connected via impedance limiting the inrush current, which is short-circuited after start-up. It is for greater performances this solution is expensive, so some authors are looking for ways to limit the number of starting impedances. Use one resistor connected to the middle conductor in combination with primitive sequential switching of individual phases. The most popular method is the sequential switching of individual phases. The sequences used are several (B–AC) with (1/4T) delay; (B–AC) with delay (2+1/4) T [2, 3]; (A–C–B) with (1/6T) delays [6] and more. For implementation, it is necessary to use separate switches in the individual phases and to know the remnant current in the transformer core. At the same time, the methods are optimized only for some group connections.

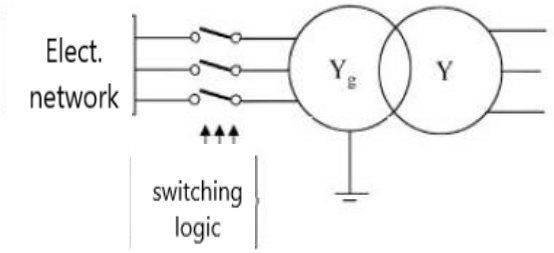

Fig.2. Simplicity of sequential switching

The same time, mechanical switches are commonly used for switching, which cannot avoid arc jumps and debouncing. In doing so, a defined pinning time must be ensured. It is therefore advisable to prefer sequences less sensitive to the accuracy of the switching times and to verify the properties of the switches. The biggest pitfall of sequential switching is detecting the initial remnant magnetization. Very often, the problematic measurement of magnetization is bypassed, for example, by a controlled sequential one by switching off the transformer [6] or by demagnetization using capacitors connected in parallel to the transformer. The magnetization measurement is then not necessary because the core is before the next pinning in a defined state.

III. USE OF UNINFORMED SOFT SWITCHING METHODS IN THREE-PHASE TRANSFORMERS

For single-phase transformers, controlled clamping is advantageously used to limit the inrush current without having to measure the initial remnant magnetization of the core. It is gradual Magnetization of the core with short pulses in individual periods, which ensure a defined magnetization of the core at the end of the sequence. We call such methods uninformed. Both methods [1,7] end at the same time sequence with a core magnetized at the saturation limit. By a suitable combination with sequential

switching from [2,3], it would thus be possible to create a configuration for limiting the inrush current of a three-phase transformer without the need to measure the initial remnant magnetization of the core. In the first phase of the solution, the quality of the sequential switching algorithm from [2,3] was verified. Two variants (B–AC) with a delay of (1/4T) shown in Fig. 3, and a similar (B–AC) with a delay were tested.

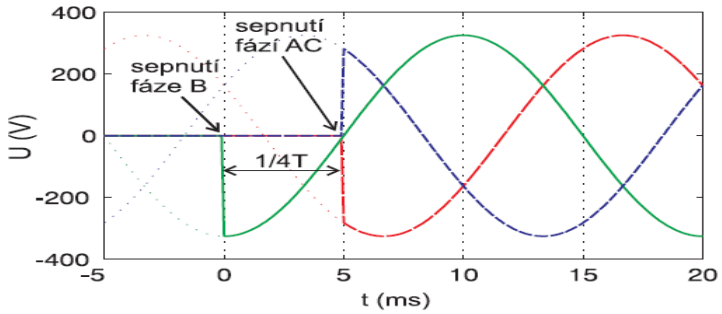

Fig.3. Tested B–AC switching sequence with 1/4T delay

For testing, an EI transformer Legrand 776862 (230//230V,10kVA) was used in the connection($Y_n y_n 0$). Connected to a star with the center drawn out. The couplings and posts are made of magnetically oriented steels, pulled bolts without interleaving, so they fit butt. So the core has considerable air gaps. The result is minimal residual magnetization. The highest measured value the initial remanent magnetization was 0.12 T. The measurement was made with the attached Hall probe at the point of contact of the coupling and the column with respect to the reference value of the amplitude of the magnetic flux in steady state operating status of the transformer. When measuring the inrush current diagrams, no initial was needed define or measure remanent magnetization. Due to this, the curves in the diagrams are only slight wavy. Fig.4,and 5 show measured diagrams for both switching sequences. Maximums are plotteof the starting current in the individual phases of the transformer L1, L2 and L3, then the starting current in middle conductor N. For clarity, the curve of the highest starting current of all phases is added in absolute value. The independent variable of the graph is the supply voltage pin-on angle measured on phase L1. For both sequences, two areas with suppressed inrush current suitable for use are evident soft start circuit. The reduced inrush current is a range of approximately (25°) of the clamping angle with the center at the maximum voltage of phase L2, which switches first. For correct operation of inrush current limitation therefore, switching must take place with an accuracy of 0.7 ms. it can be seen from the diagrams that both methods have a similar effect. The second variant with a delay (3 + 1/4) T shows a smaller inrush current when switched at an inappropriate angle, which is not for the application substantially. It is therefore more advantageous to use the faster first variant.

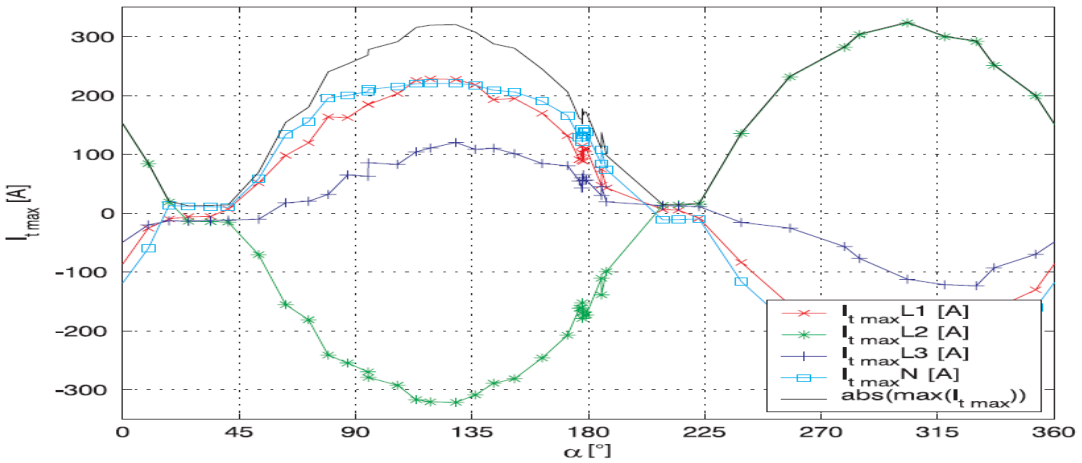

Fig.4.Diagram of maximum inrush current during sequential B–AC switching with 1/4T delay Files: In3fMode2_R50load_ProhozenoT1aT2_xx.

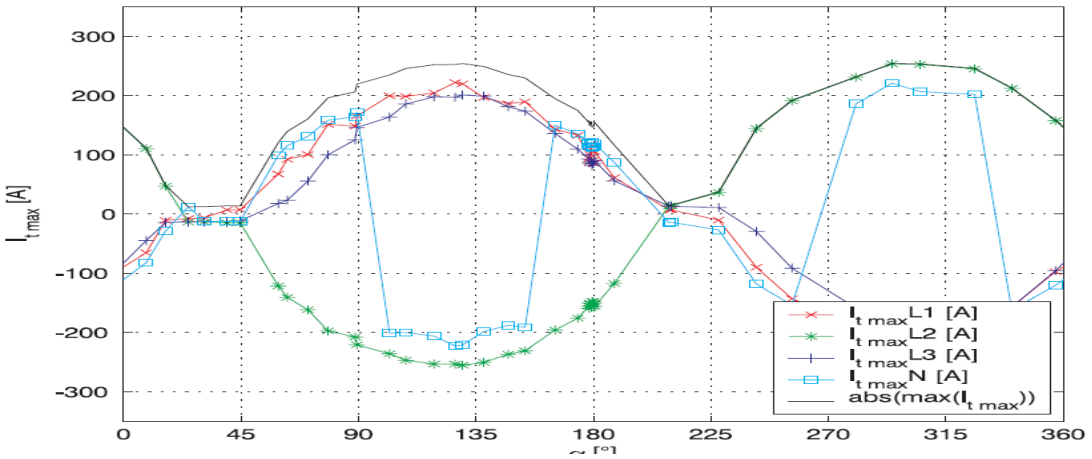

Fig.5. Diagram of maximum inrush current during sequential B–AC switching with delay (3+1/4) T
Files: In3fMode4_R50load_ProhozenoT1aT2_xx

IV. CONCLUSION

The combination of uninformed methods of limiting the inrush current and sequential switching is a suitable, technically simple, solution for three-phase transformers of smaller powers. For large machines, it is more advantageous to measure the initial remanent magnetization compared to the cost of the semiconductor switching element.

References

[1] Novák M. Přechodový děj při zapnutí transformátoru – způsoby omezování zapínacího proudu. [on-line] [disertační práce] Liberec: TU, 2003. 256 s. [cit. 29. 9. 2005].

[2] Brunke J. H., Fröhlich K. J. Elimination of Transformer Inrush Currents by Controlled Switching. Part I - Theoretical Considerations. [on-line] IEEE Transaction on Power Delivery, vol. 16, No. 2, April 2001 [cit. 29. 9. 2005] .

[3] Brunke J. H., Fröhlich K. J. Elimination of Transformer Inrush Currents by Controlled Switching. Part II – Application and Performance Considerations. [on-line] IEEE Transaction on Power Delivery, vol. 16, No. 2, April 2001 [cit. 29. 9. 2005] .

[4] Yu Cui, Abdulsalam S. G. A Sequential Phase Energization Technique for Transformer Inrush Current Reduction — Part I: Simulation and Experimental Results.[on-line] IEEE Transactions on 36 EPVE 2005 power delivery, vol. 20, No. 2, April 2005. pg. 943–949 [cit. 29. 9. 2005].

[5] Yu Cui, Abdulsalam S. G. A Sequential Phase Energization Technique for Transformer Inrush Current Reduction — Part II: Theoretical Analysis and Design Guide.[on-line] IEEE Transactions on power delivery, vol. 20, No. 2, April 2005. pg. 950–957 [cit. 29. 9. 2005] .

[6] Prikler L., Bánfai G., Bán G., Becker P. Reducing the magnetizing inrush current by means of controlled energization and de-energization of large power transformers. [on-line] International Conference on Power System Transients, IPST 2003, New Orleans [cit. 30. 9. 2005].

[7] Fraunhofer-Gesellschaft Zur Förderung der Angewandten Forschung E. V. Procedure and equipment for avoiding inrush currents. Erfinder: Konstanzer, Mlohael. European Patent Office. Patentschrift, EP0 575 715 B1. 1993-12-29.

[8] A.Honzák, P.Nykodým, J.Koláčný: Bifurcation analysis of DC drive. Proceedings XXVIII. National Conference of Electric Drives - Pilsen 2003, pp.223 - 228 .

[9] Adithya Ballaji, Nagaraj Hediyal, Dr. Rajashekar P Mand, K Narayana, "Design and Implementation of Perturb and Observation Maximum Power point Transfer (MPPT) algorithm for Photovoltaic system" Adithya Ballaji Int. Journal of Engineering Research and Application Vol.8 2016..

International Journal of Innovation Engineering and Science Research
(ISSN: 2581-4591)
www.ijiesr.com

Problems of Reading and Transliterating Urdu into Hindi script: An Approach to Human and Machine Transliteration

Dr. Sabahuddin Ahmad
Department of Linguistics
Aligarh Muslim University, Aligarh

ABSTRACT

This paper briefly examines the orthographic structure of Urdu and difficulties involved in reading and transliterating Urdu script by both-human and machine. The Alphabetical inventory of Urdu represents only four vowel graphemes out of total 36 basic characters i.e. [ا،و،ی،ے]. Rest of the vowel sounds are represented by the diacritical marks.
These vowel graphemes are pronounced differently in different position and most of the time these graphemes depend upon the diacritic marks to represent the value of different vowels.

Most of the vowels are represented with the help of diacritic marks which are usually not maintained in regular writing. Absence of these diacritic marks poses lots of problems and error for both human and machine. In Devanagari, there is one to one correspondence between grapheme and the sound. Hindi speakers learning Urdu script face lot of problems in reading and writing Urdu script. The absence of diacritic marks in input text of Urdu also causes errors in the Devanagari output of the automatic transliteration system. (Malik et. al 2009) have shown in their research that the Hindi Word Language Model increases the accuracy of Urdu Hindi transliteration especially for Urdu input without diacritical marks but diacritical marks are crucial and necessary for Urdu to Devanagari transliteration.

Apart from this, a consonant phoneme is realized by a number of graphemes or combination of two graphemes in Urdu. These irregularities cause problems in reading and transliterating Urdu script. Empty graphemes of Urdu script also create problems in transliterating Urdu script into Devanagari script. This paper will explore and explain these problems of Urdu script with special reference to Devanagari.

Key Words: Transliteration, Graphemes, Consonant, Vowel, Diacritical Marks

I. Introduction

Urdu script, known as ‘nastaliq’ writing, is an adoption of Perso-Arabic script with certain addition and modification. The alphabetic inventory of Urdu, represent 36 basic characters (for some it is 35). Out of these 36 graphemes, there are only four characters/graphemes for vowel sounds, as shown in the table no. 1 below:

Urdu Vowel Graphemes	Name
ا	əlɪf
و	wao
ی	çoʈi: je
ے	bəɽi: je

Urdu hasalso diacritical marks which are either used alone or in combination with the above Urdu vowel graphemes to represent vowel sounds in Urdu script. These diacritical marks are shown in the table no. 2 below:

Urdu Vowel Diacritics	Name of the Graphemes	Vowel Sounds
◌َ	zəbər	/ə/
◌ِ	zer	/ɪ/
◌ُ	peʃ	/ʊ/
◌ٗ	ʊlʈa: peʃ	/u/
◌ٓ	məd	-

'◌ٓ' /məd/ does not occur alone to represent any sound. It is used above 'ا' /əlɪf/ to represent /a:/ sound as in the word 'آم'/a:m/ 'mango'.

The vowel grapheme 'ا' /əlɪf/ in the initial position of the word works as an entity which has no sound in itself; it requires diacritic marks to represent any vowel sound.

I. [ا] 'Alif'+Diacritic: [ا] 'Alif' at initial position with different diacritics represents different vowel sounds:

(a) Lax close unrounded front vowel

[ا] with [◌ِ]= [اِ] - [ɟ]- /ɪ/ 'اِملی' /ɪmlɪ/

(b) Lax half-open unrounded mid vowel

[ا] with [◌َ]= [اَ] - [ɜ] - /ə/ 'اَنداز' /əndɑ:z/

(c) Lax close rounded back vowel

[ا] with [◌ُ]= [اُ] - [ɔ̈]- /ʊ / 'اُلو' /ʊllʊ/

(d) Tense open unrounded back vowel

[ا] with [~]= [ا̃] - [ɜ̃]- /ɑ/ 'آم' /ɑ:m/

Apart from the diacritics, [ا] 'Alif' at initial position also needs other vowel graphemes to represent tense close unrounded front vowel /i/ and tense close rounded back vowel /u/ sound. Without these vowel graphemes these sounds cannot be represented:

II. [ا] 'Alif'+ Another Vowel Grapheme+Diacritic: There are some vowels sounds which need three Urdu characters for their representation in Urdu script.

(a) Tense close unrounded front vowel : /i/

[ا]+[ی] +[◌ٕ]= [ɨ̟] - /i/ 'اِیجاد' /i:dʒa:d/

(b) Tense close rounded back vowel: /u/

[ا]+[و] +[◌ُ]= [ɔ̝] - /u/ 'اُوپر' /upər/

The absence of diacritics from the above letters will change their nature and will be read as /e/ and /o/ instead of /i/ and /u/.

III. [ا] 'Alif'+ Another Vowel Grapheme: Tense half-close unrounded front vowel /e/ and tense half-open unrounded front vowel/o/ are represented by combining two vowel graphemes of Urdu. They do not need any diacritic to represent vowel:

(a) [ا]+[ے]= [ए] - /e/ 'اے' /e/

For example- /ek/ 'ایک', 'एक'

(b) [ا]+[و]= [ओ] - /o/ 'او' /o/

For example- /or/ 'اور', 'ओर'

II. Positional Variation in the form of Vowel Characters

One character i.e. [ا] is common in all the vowel characters discussed in the previous section but the occurrence of this [ا] is limited to the initial position of the word. In medial and final position of the word, [ا] does not occur with the diacritics and other vowel characters which involve in representing particular vowel sound. It is similar like matras e.g. [ी, ु etc] of Hindi, which represent vowel sounds in middle and final position of the word. Consider following examples of Table 1:

Table 1

Devanagari Characters	Perso-Arabic Characters	IPA	Position	Examples
ई–ع	ایـ	/i/	I	ایمان 'ईमान' /ɪman/
ई	ئ	/i/	F	کئ 'कई' /kai/
ी	ی	/i/	M	پیلا 'पीला' /pila/
ी	یِ	/i/	F	کہِی 'कही' /kəhɪ/
उ	اُ	/ʊ/	I	اُس '' /ʊs/
ु	ُ	/ʊ/	M	کُچھ '' /kʊc^{h}/
ऊ	اُو	/u/	I	اُوپر '' /upər/
ू	ُو	/u/	M	جُون '' /dʒun/
ू	ُو	/u/	F	بھالُو '' /b^{h}alu/

Urdu script has more variations in its character representation than the Devanagari script, as shown in table 1. In Urdu, /i/ vowel is represented by four different sets of characters depending upon its position of occurrence whereas in Devanagari, it is represented by two characters only i.e. 'ई' and 'ी'. Similarly other vowels have different forms of character sets according to the position. This is a tricky thing in itself to learn the reading and writing of the Urdu script, though this is not the problem at all for

automatic transliteration now. An accurate mapping of the characters provides good output of the transliteration but still there are some features and characteristics of Urdu script which poses problems for both human and machine.

III. Features Posing Problems in Reading and Transliteration Urdu Script

Absence of diacritic marks from the script, multiple consonant characters for single sound, empty letters/grapheme, characters for special Urdu sounds are the major reasons for the complexity of the Urdu script.

3.1. Absence of Diacritic Marks

The diacritical marks, which play a very significant role in Urdu phonology, however, remain unmarked in regular text. Native speaker of Urdu realize them correctly but it is difficult for the non-native speakers or foreign learners. Without the diacritic marks words seem ambiguous to read and sometimes it is unpredictable. This is challenging for both human and machine. Consider the following examples of output taken from SANGAM- automatic transliteration tool (https://sangam.learnpunjabi.org/) developed by Punjabi University, accuracy of which is above 90%:

1. کیا اس کے پاس **جو** ہے۔ تمہارے پاس کتنے کلو **جو** ہے

IPA: kya ʊske pas dʒəɔ hɛ. tʊmhare pas kɪtne kɪlo dʒəɔ hɛ

ST[1]: क्या**<u>उसके</u>**पास**<u>जो</u>**है।तुम्हारेपासकितनेकिलो**<u>जो</u>**है।

IPA: kya ʊske pas **<u>dʒo</u>** hɛ. tʊmhare pas kɪtne kɪlo **<u>dʒo</u>** hɛ

OT: क्याउसके/इसकेपासजौहै।तुम्हारेपासकितनेकिलोजौहै?

"Does he has **barleys**. How many kilo **barleys** do you have?"

2. کیا یہ الّو ہے؟

IPA: kja yeh ʊllʊ hɛʔ

ST: क्यायेअल्ोवहै

OT[2]: क्यायहउल्लूहै?

"is it an owl?"

In (1) and (2), we can see that two words i.e. 'جو' /dʒəɔ/ and 'الّو' /ʊllʊ/ are transliterated as 'जो' /dʒo/ and 'अल्ोव' /əlov/ respectively which are incorrect. Correct transliterations of these words are 'जौ' /dʒaɔ/ and 'उल्लू' /ʊllʊ/ respectively. It is only because of the absence of the diacritic marks (AIRAAB) for vowels.

[1]ST- SANGAM Transliterator
[2] OT- Original Transliteration

Some more result of transliteration by SANGAM

Input Text:

انٹیا کے وزیراعظم نریندر مودی نے کہا ہے کہ انٹیا کے لیے اگلے25 سال انتہائی اہم ہیں، اس میں سب کو خود سے پانچ وعدے کرنے اور انہیں پورا کرنے کی ضرورت ہے۔

این ڈی ٹی وی کے مطابق ملک کے 75 ویں یوم آزادی کے موقع پر قوم سے خطاب کرتے ہوئے نریندر مودی نے بتایا کہ کیسے 2047 تک ملک کو اسے بنانے والوں کی خوابوں کے مطابق ڈھالا جا سکتا ہے۔

انہوں نے ایک ایک کر کے پانچ وعدے گنوائے اور زور دیا کہ سب ان پر عمل کریں اور یہ عزم بھی دوہرایا کہ وہ خود بھی ان پر عمل کریں گے۔

Output Text:

इंडियाकेवज़ीर-ए-आज़मनरेंद्रमोदीनेकहाहैकिइंडियाकेलिएअगले25 सालइंतिहाईअहमहैं,
इसमेंसबकोख़ुदसेपाँचवाअदेकरनेऔरउन्हेंपूराकरनेकीज़रूरतहै

इनडीटीवीकेमुताबिक़मलिककेे75 वींयौम-ए-आज़ादीकेमौक़ापरक़ौमसेख़िताबकरतेहुएनरेंद्रमोदीनेबतायाकिकैसे2047 तकमलिककोउसेबनानेवालोंकीख़ाबोंकेमुताबिक़ढालाजासकताहै

उन्होंनेएकएककरकेपाँचवाअदेगँवाएऔरज़ोरदियाकिसबउनपरअमलकरेंऔरयेअज़मभीदोहरायाकिवोख़ुदभीउनपरअमलकरें गे

Table 2: Result Analysis

S.No.	Input Words	IPA	No. of Occurrence in text	Wrong Outputs	Correct words	Reasons
1	وعدے	/vade/	2	वाअदे	वादे	wrong mapping/special character
2	این ڈی ٹی وی	/en di ti vi/	1	इनडीटीवी	एनडीटीवी	wrong mapping
3	موقع	/maɔqa/	1	मौक़ा	मौक़े	different convention of reading
4	ملک	/mʊlk/	2	मलिक	मुल्क	absence of diacritics (AIRAAB)
5	خوابوں	/xwabõ/	1	ख़ाबों	ख़्वाबों	wrong mapping
6	گنوائے	/ginvae/	1	गँवाए	गिनवाए	absence of diacritics (AIRAAB)

We can see in the table 2 that the words of S. No. (4) and (6) i.e. 'ملک' 'country' / mʊlk/ and 'گنوائے' 'to make someone count' /ginvae/ respectively are wrongly transliterated as 'मलिक' 'name or title of the name' /məlɪk/ and 'गँवाए' 'lost' /gãvae/ due to the absence of diacritic marks. Such errors affect the meaning of the text i.e. sometimes such errors totally change the meaning of a text, causes wrong information and misunderstanding.

3.2 Multiple Consonant Characters for One Sound

Like vowels, consonant sounds also have more than one character or graphemes to represent single sound. In other words, single sound has multiple correspondences in Urdu script but they are not positionally governed. These multiple characters have random distribution.

Table 3

Consonants	IPA	Hindi Alphabets	Urdu Alphabets			
Voiceless Dental Stop	/t/	त	ت	ط		
Voiceless Alveolar Fricative	/s/	स	س	ص	ث	
	/h/	ह	ہ	ح		
Voiced Alveolar Fricative	/z/	ज़	ض	ذ	ز	ظ

Urdu has many loan words from Arabic and Persian that include graphemes from these languages, retained in the Urdu spelling. As a result, there are several different Urdu characters mapping to the same phoneme. These graphemes which represent same sound are of different origin so that it is hard to infer any pattern in their distribution. It causes errors in the spelling of words.

Table 4

Urdu Graphemes	IPA	Urdu Words	Words in Devanagari	IPA	Meaning
ت	/t/	تماشہ	तमाशा	/təmaʃa/	show
ط	/t/	طشتری	तश्तरी	/təʃtəri/	plate
س	/s/	سانس	सांस	/sãs/	breath
ص	/s/	صبر	सब्र	/səbr/	patient
ث	/s/	ثمر	समर	/səmər/	fruit
ح	/h/	حلوہ	हलवा	/həlwa/	sweet dish
ہ	/h/	ہلچل	हलचल	/həlcəl/	stir or bustle
ض	/z/	مضمون	मज़मून	/məzmun/	essay
ذ	/z/	ذرّہ	ज़र्रा	/zərra/	particle
ز	/z/	مزدور	मज़दूर	/məzdur/	labour
ظ	/z/	ظالم	ज़ालिम	/zalɪm/	cruel

We can see in the table 4 that the identification of correct graphemes which has multiple correspondences is a real challenge while transliterating the Devanagari text into Urdu text. Similarly, it is confusing for the Urdu learners and sometimes for native speakers too. Multiple characters are the major challenge for Devanagari to Urdu automatic transliterator. For the learners, practice is the only tactics to come out this difficulty.

3.3 Empty letters/graphemes

Urdu owing to its Perso-Arabic heritage displays an interesting phenomenon of silence of graphemes where either an individual grapheme or even a sequence of graphemes is silent in Urdu orthography. It means that the sounds of these silent characters are not transferred in the target script.

Table 5

Urdu Words	IPA (Pronunciation)	Letters forming the word	Silent sequence of Grapheme	Automatic Transliteration	Correct Transliteration
على الصباح	/əl:ssəbah/	ع، ل، ی، ا، ل ص، ب، ا، ح	ل، ی، ا	अलस्सबाह	अल्स्सबाह
شمش الحق	/ʃəmʃʊlhʊda/	ش، م،ش، ا،ل، ہ د، ا	ا	शमशउल-हक़	शमशुलहक़
قمرالزماں	/q:mrʊzzəmã/	ق، م، ر، ا، ل، ز، م، ا، ں	ا، ل	क़मरअल्ज़मां	क़मरूज़्ज़मान
صباح الدین	/səbahʊddin/	ص، ب، ا، ح، ا، ل، د، ی، ن	ا، ل	सबाहउद्दीन	सबाहुद्दीन
انالحق	/ənəhəq/	ا، ن، ا، ل، ح، ق	ا	अनालहक	अनलहक
قمرالہدا	/qəmrʊlhʊda/	ق، م، ر، ا، ل، ہ د، ا	ا	क़मरअलहदा	क़मरुलहुदा
قمرُالہدا	/qəmrʊlhʊda/	ق، م، ر، ُ، ا، ل، ہ د، ا	ا	क़मरुअलहदा	क़मरुलहुदा

3.4 Special Characters

Perso-Arabic Letter	Devanagari	Roman Script
ع	--	**A**
ء	--	**'a**

Mehwish Leghari and Mutee U Rahman Arain, 2015

We can represent these two sounds by the character representing above sound in Roman script by making minor changes using diacritic marks but we don't find any mean to represent them in Devanagari. They are alphabets, not diacritic marks but they don't represent any sound alone.

Alphabets representing Common phonemes in Hindi-Urdu		Alphabets representing special Phonemes of Urdu	
ک	क	ق	क़
کھ	ख	خ	ख़
گ	ग	غ	ग़
پھ	फ	ف	फ़
ج	ज	ز﴿ذ، ض﴾ظ	ज़

Apart from multiple character we should also focus on the special alphabets representing special sound which is the beauty of the Urdu script

IV. Conclusion

To sum up, it can be said that the absence of 'Airaab' from the Urdu script poses major problems of ambiguity in reading Urdu script and transliterating Urdu script into another script e.g. Devanagari. Empty letters i.e. the letters which are not supposed to be pronounced as per the convention of the Urdu script are another source of error while reading Urdu text. They create problems in automatic transliteration as well. Feature of 'multiple characters for one sound' may not be much problematic in reading text but creates lots of confusion while writing in Urdu or transliterating any text into Urdu. This paper has highlighted the complexities of the Urdu script and discussed the problems which occur in reading Urdu script and the problems occur in automatic transliteration of Urdu script into any other script.

References

[1] Bushra J. and Tafseer A. (2009) "Hindi to Urdu Conversion: Beyond Simple Transliteration". *Proceedings of the Conference on Language & Technology*, Lahore, pp 24-31.

[2] Malik, A. et.al. (2009). "A Hybrid Model for Urdu Hindi Transliteration" in the Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pp. 177-185, Suntec, Singapore

[3] Ahmad, Sabahuudin. (2009). *UrduHarf-o-Saut*.2010.Delhi:Arshia Publication,

[4] _____ (2014). *Imlanama :Takhreeb vo radde takhreeb*. Delhi: ArshiaPublication,

[5] Yamunà Kachru. Hindi. John Benjamins Publishing Co., (2006)

Estimation of Ground Water Potentials in Bauchi Metropolis, Bauchi State, Nigeria.

[1]Lagasi, J. E; [2]Mije, F. G; [3]Abdulrahman. B; [4]Basu. M. M; [5]Bello. A.

[1]*Department of Civil Engineering, University of Jos.*
[2]*Department of Civil Engineering, University of Jos.*
[3]*Department of Civil Engineering, University of Jos.*
[4]*Department of Mining Engineering, University of Jos.*
[5]*Water and Sanitation WASTAN Agency, Bauchi.*

ABSTRACT

This study aimed at estimating underground water potential in Bauchi Metropolis a case study of Gwallameji, behind NITEL Office, opposite emergency ward, Old GRA opposite Niima and Shafa Gidan Mai). Rainfall, temperature and humiditydata of the study area were collected from the meteorological center, Bauchi, pumping test were conducted on the existing Boreholes.Also, geophysical survey on the existing Boreholes to determine their Resistivity Values was carried out. The pumping test was conducted using constant head, while geophysical survey was conducted using Schlumberger array and five vertical electrical sounding (VES) points were sounded. The draw down and their corresponding residual draw down are plotted respectively.Data collected are interpreted using Win Resist 2 software. The results show that the rainfall pattern couple with temperature and relative humidity record contributes to the ground water storage, the average depth to water table in the crystalline basement in the study area is 4.2m and the average yield of 0.87 l/sec. (52.2 l/min.), This was determined from the results obtained in the area. The average transmissibility in the study area is 7.90X10-3 m2/sec with the resistivity value ranges from 24.8 to 847.1 ohms-meter and their permeability are about moderate to low and the storability in the basement complex is reasonable; it ranges from 1.002x10-3 to 1.008x10-3. The findings clearly reveal the reliability of ground water potential in the metropolis. This investigation should be extended to other areas in Bauchi hitherto not covered in this work.

Key words: Underground, borehole, resistivity, basement, permeability.

I. INTRODUCTION.

Groundwater is the main source of water for much of the rural population of sub-Saharan Africa, particularly those in the drought-prone semi-arid regions Okereke (1997). Groundwater is a fundamental component of the water resources for domestic, industries and drinking purpose. The exploitation and sustainability of this resource is key to human survival and economic development. Large areas of sub-Saharan Africa are underlain by crystalline basement rocks and much research has been undertaken on groundwater storage and flow within this aquifer and how best to develop and protect water supplies (Emenike, 2007).

The development of groundwater in the study area is beset with problems of failed (Abortive) boreholes. The groundwater in the basement complex terrain is mainly contained in the porous and

permeable weathered zones. The groundwater yield from the weathered horizon is often supplemented by the accumulated groundwater in the fractured and jointed column of the Basement complex rocks(Okolie, 1999).

The hydrogeological and hydraulic investigation involves monitoring the static water level, depth of well/borehole and calculation of aquifer parameters such as Transmissibility, Specific capacity, hydraulic conductivity, Storability which were used to evolve a groundwater potential map of the study area (Leduc 2009).

The properties of water make it suitable for human beings to survive in differing weather conditions. Water is characterized by complex anomalous properties that differentiate it from other substances. Water is the universal solvent due to its polar nature. It dissolves a large number of different chemical substances (Shamang, 2005).

Despite the long history of groundwater extraction throughout civilization, it has only been during recent decades that the use of groundwater has grown exponentially. Most people depend on groundwater for their daily supplyOyewoye (1963). Although every society has its own problems and Bauchi is not an exception yet the very recent challenges of ground water in various locations have been a challenge for the government and the people of Bauchi state.

Even with the supply of water to people of the metropolis by the State Water Board through the Gubi Dam treatment plant, yet other locations where reticulation could not reach suffer inadequate water supply due to abortive wells/boreholes. This necessitated this investigation which will provide viable solution to the water need of the study area.

II. Methodology.

Hydrogeological map of the study area was obtained showing the bore hole locations in the study area.The map reveals the fractured areas.

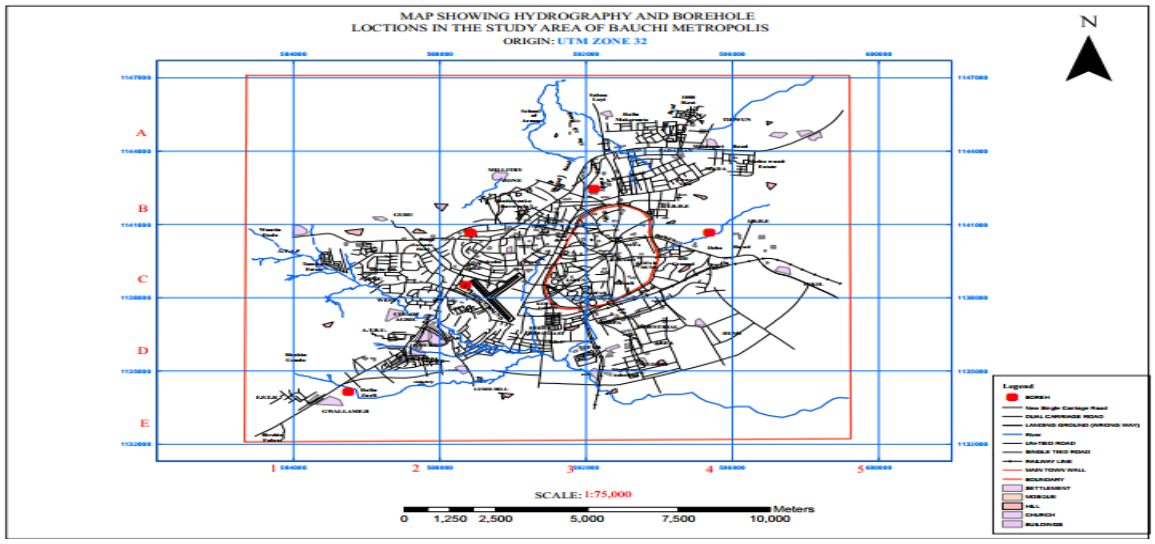

Fig 1: Showing bore hole locations in Bauchi.

The Rainfall Data for eleven (11) years between 2011 to 2021 were collected from the meteorological center, Old Airport in Bauchi, Bauchi State of Nigeria and analyzed. The meteorological elements of concern include Rainfall, Minimum Temperature., Maximum Temperature and Humidity.

The electrical resistivity method is the single the most important method used in hydro- geological investigation and it is useful because the resistivity of a formation is largely dependent upon its moisture content. The Schlumberger configuration or arrangement of electrical resistivity was then used.The survey is based on evaluating the apparent resistivity Ra**.**

Electrical Resistivity prospecting method has been used in determining groundwater potential Areas of Gwallameji, Old GRA, Behind Nitel office, Mudalawal Market, Gidan Mai of Bauchi Metropolis, Bauchi State, using the Vertical Electrical Sounding (VES) method at different selected points within the Metropolis. A total of five (5) VES points were investigated in the study areas. Schlumberger array with Resistivity meter (McOHM-EL Model 2119) was used in the Geophysical survey. The Data was interpreted with the software called WinResist2.

The pumping test data was analyzed byCooper-Jacob's Method wherethe modified non-steady state formula was applied thus:

$$S = \frac{2.25Tt0}{r^2 S} \quad \text{---------- (1)}$$

It follows that transmissibility:

$$T = \frac{2.30\ Q}{4\pi\Delta_s} \quad \text{---------- (2)}$$

III. Results and Discussion.

3.1. Results Presentation.

Bauchi Metropolis entirely is controlled climatically by two main factors, namely altitude and the position across the path of the season migration of Intertropical Convergence Zone (ITCZ). These causes seasonal alternation depending on which of them is dominant. As a result of this, there is a period of wet season alternation with very dry season in the study area, classifying it as wet dry tropical climate region. The wet season starts in late April and ends in early October with the highest rainfall in August, while the dry season is from late October to early April.

As shown in figure1, the rainy season mostly startsfrom the month of April and ends by September/October of every year depending on its pattern. July and August have the highest mean monthly rainfall of 187.67mm and 250.53mm, while April and October have the lowest of 0.72mm and 18.76mm.

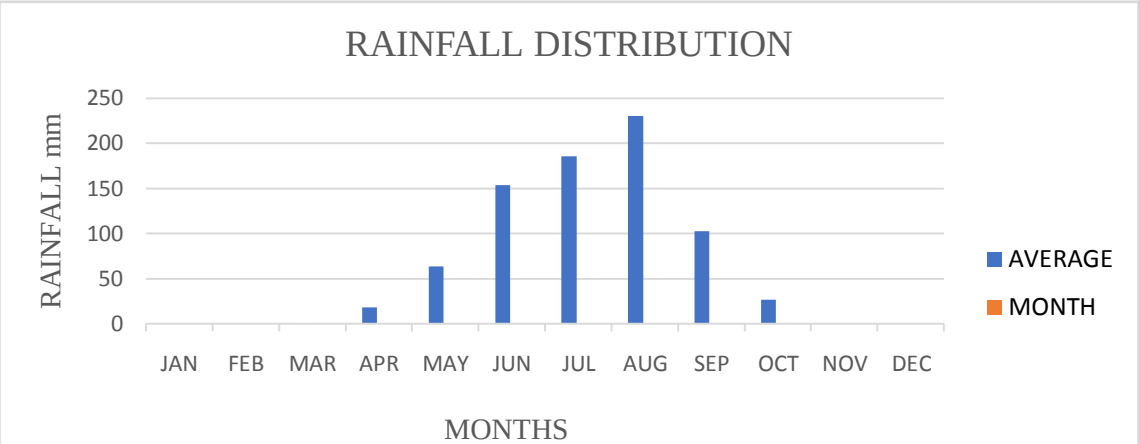

FIG.2: AVERAGE MONTHLY RAINFALL OF THE STUDY…

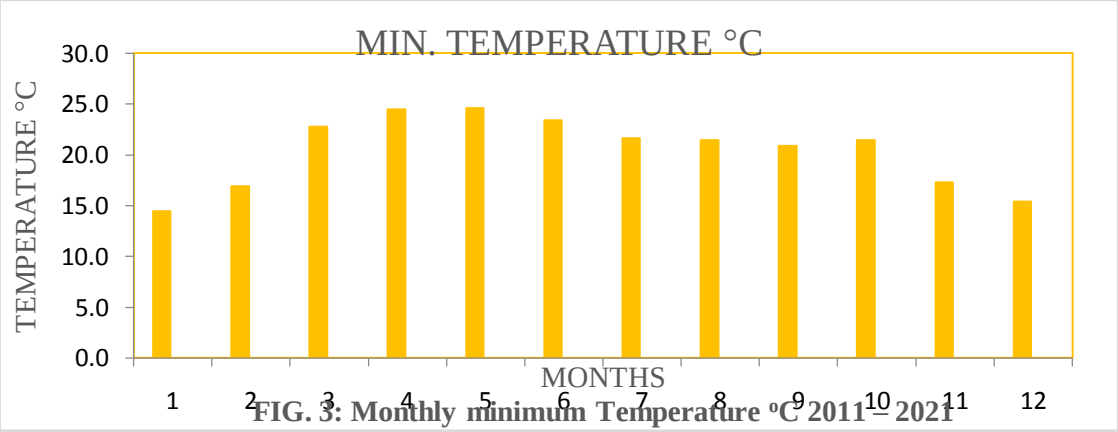

FIG. 3: Monthly minimum Temperature °C 2011 – 2021

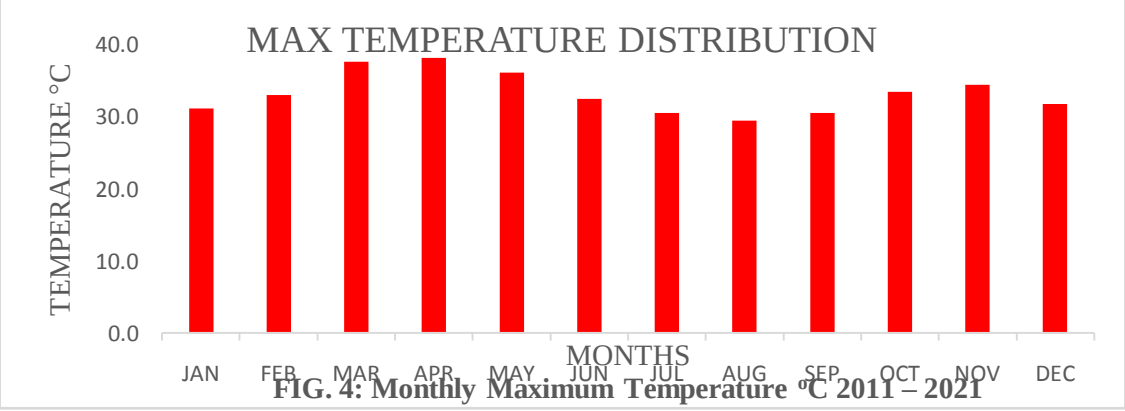

FIG. 4: Monthly Maximum Temperature °C 2011 – 2021

There are two distinct seasons; hot and dry season with north easterly winds between February and May, cool and wet season between June and September with southwesterly winds and cool dry dusty and windy harmattan season from October to January of each year.
Bauchi is hottest in the month of April with average temperature of 38.1°C, while the coldest months are December and January, when temperature falls to 14.4°C

The mean monthly relative humility for 2011-2021 is shown in figure 4, indicating high levels of relative humidity in the month of July of about 87.8%. It drops to 26.7% in the harmattan season.

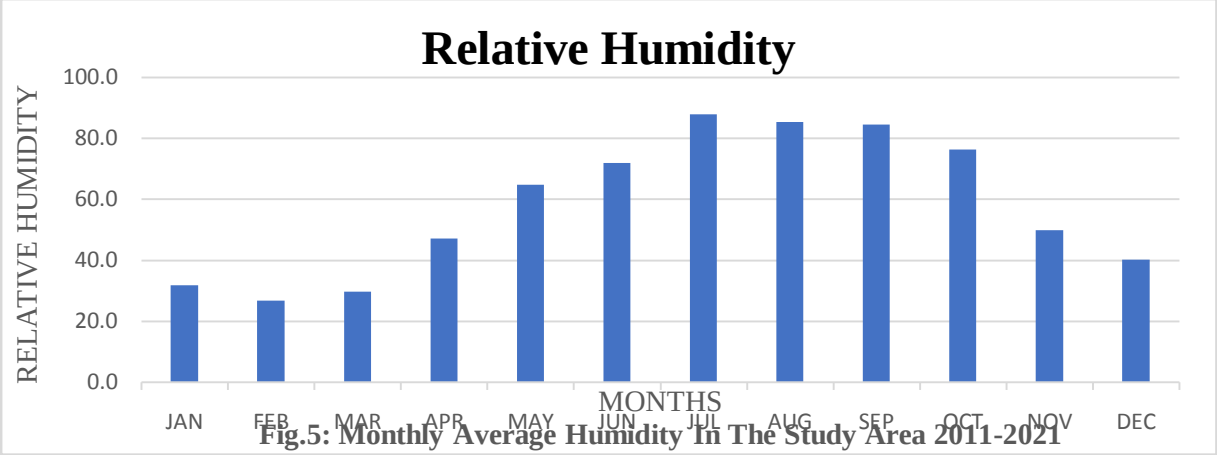

Fig.5: Monthly Average Humidity In The Study Area 2011-2021

The boreholes in the study areas were subjected to pumping test, and their draw down are plotted against time with their corresponding residual draw down from figure 6 -15 to show their respective response to pumping.

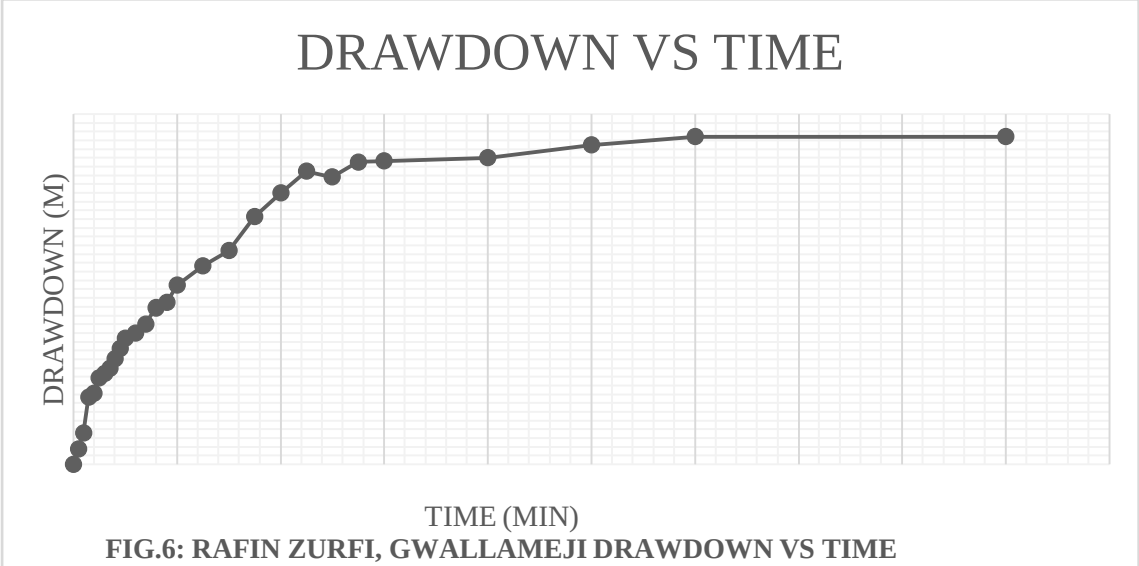

FIG.6: RAFIN ZURFI, GWALLAMEJI DRAWDOWN VS TIME

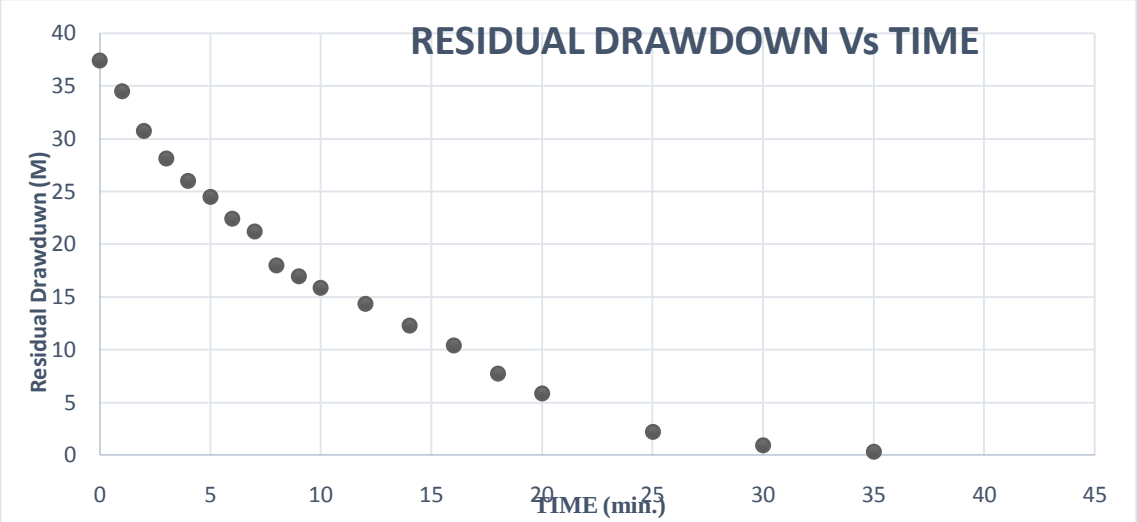

Fig.7: RAFIN ZURFI, GWALLAMEJI RESIDUAL DRAWDOWN Vs TIME

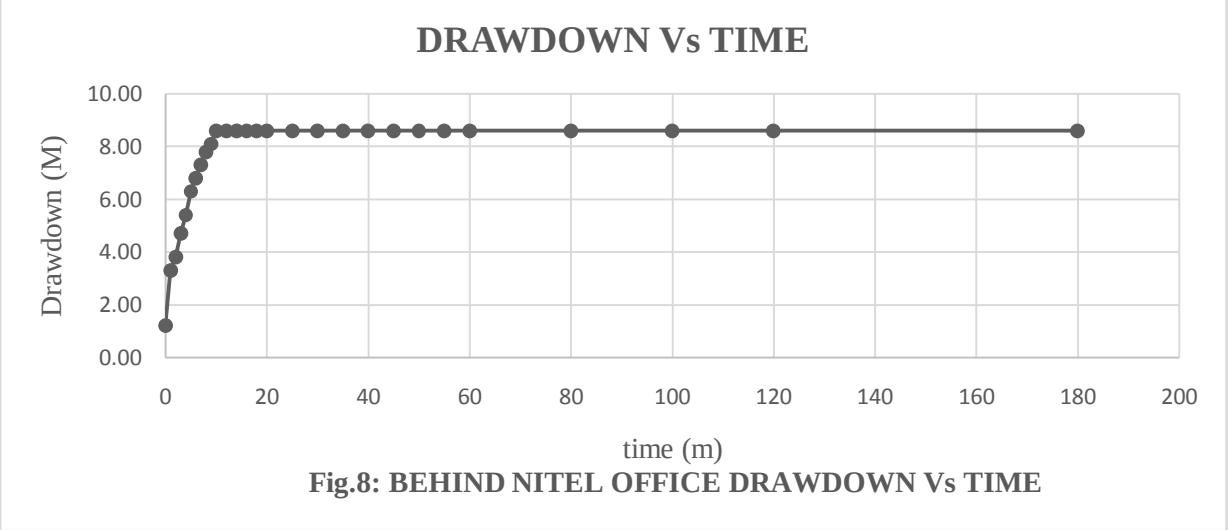

Fig.8: BEHIND NITEL OFFICE DRAWDOWN Vs TIME

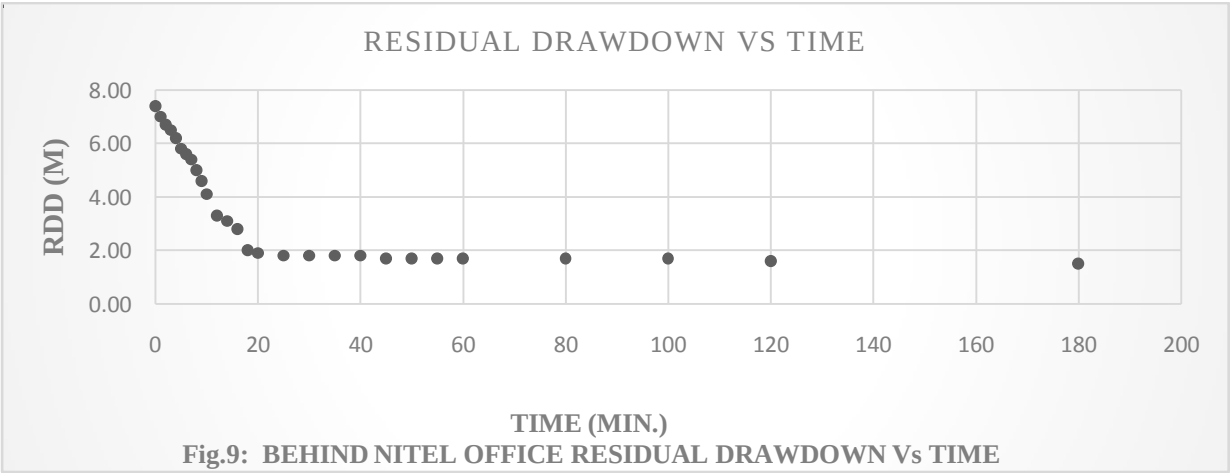

Fig.9: BEHIND NITEL OFFICE RESIDUAL DRAWDOWN Vs TIME

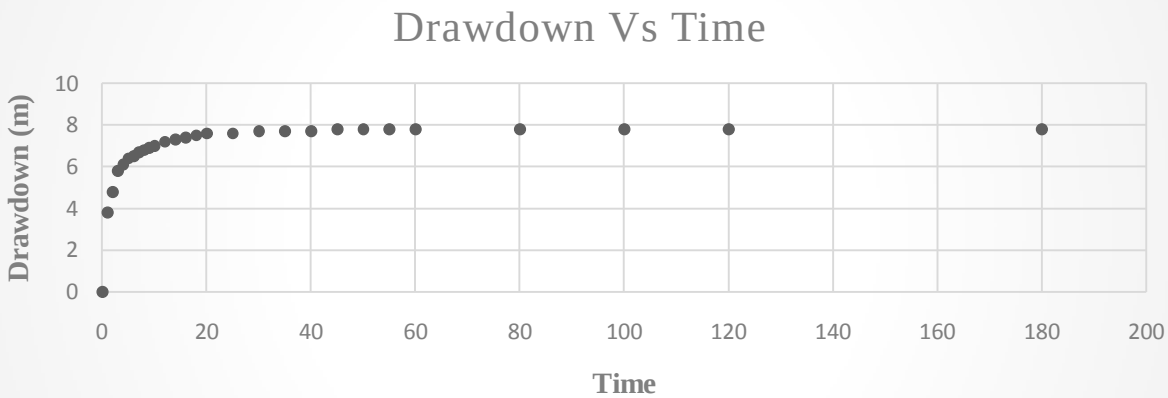

Fig.10: OPP. EMERGENCY DRAWDOWN Vs TIME

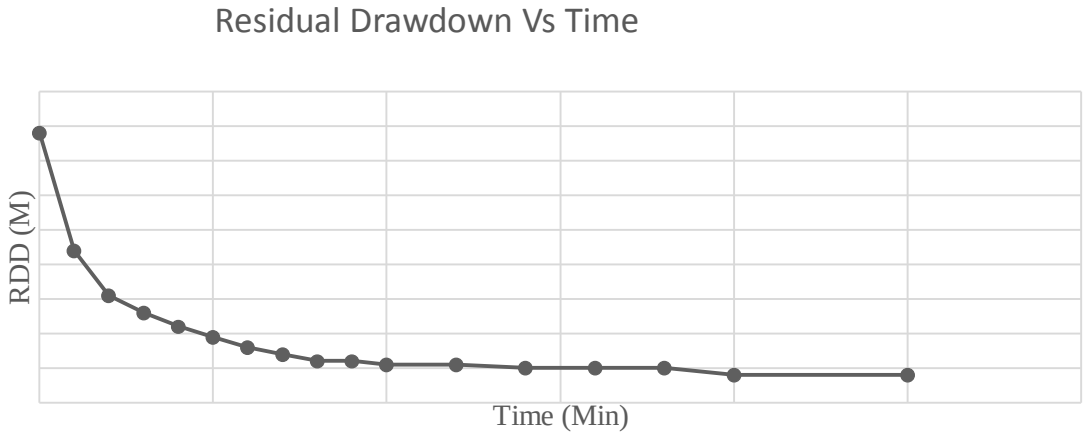

Fig.11: OPP. EMERGENCY RESIDUAL DRAWDOWN Vs TIME

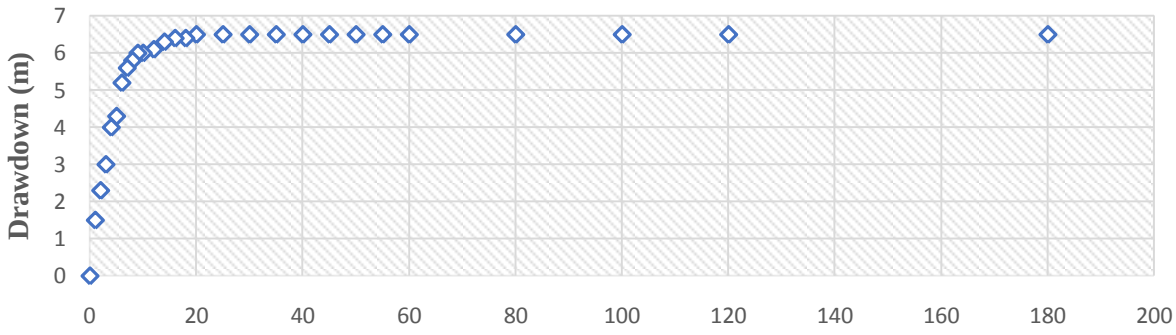

Fig.12: OPP. NI'IMA DRAWDOWN Vs TIME

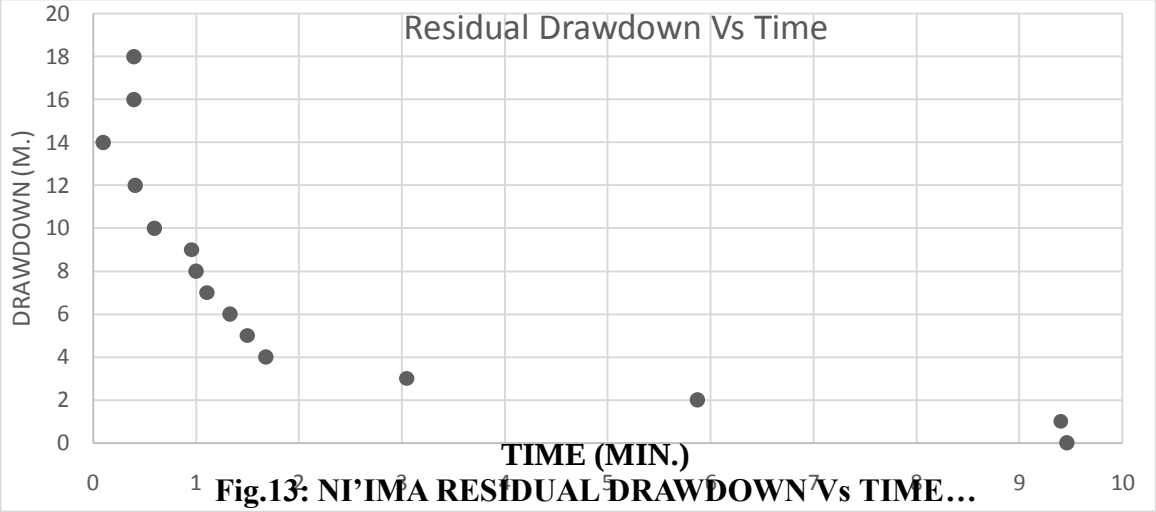

Fig.13: NI'IMA RESIDUAL DRAWDOWN Vs TIME…

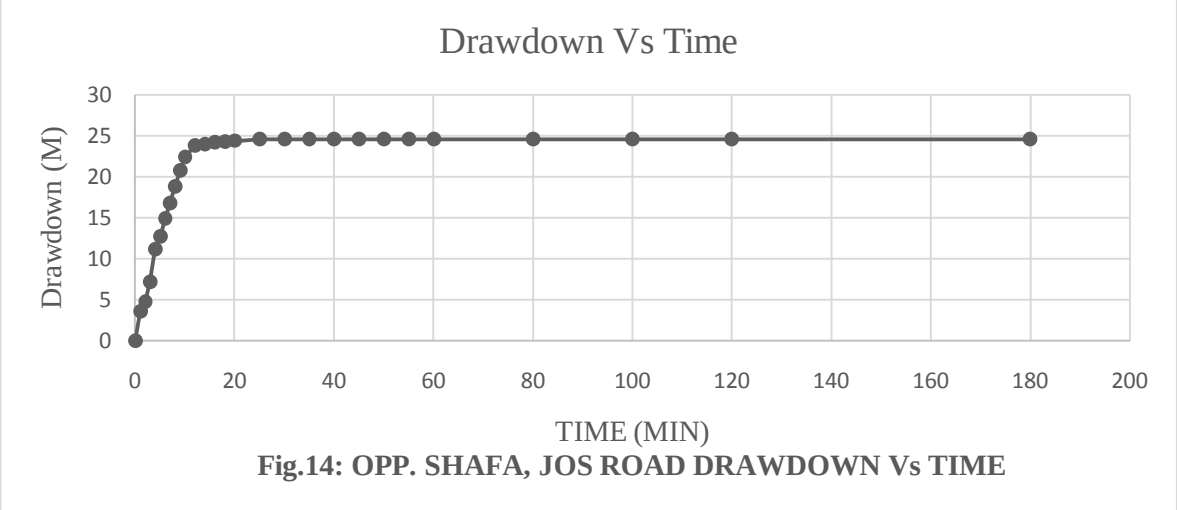

Fig.14: OPP. SHAFA, JOS ROAD DRAWDOWN Vs TIME

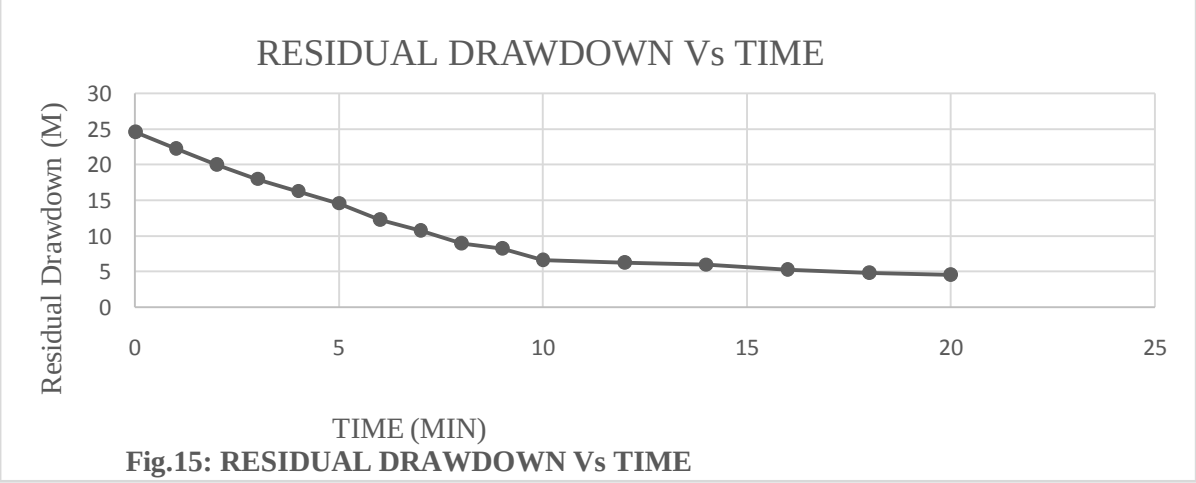

Fig.15: RESIDUAL DRAWDOWN Vs TIME

Boreholes are completed in moderately and poorly decomposed gneisses and granites and in fractured pegmatite, aplite and quartz veins, with general depths of20 to 50 meters and yields 10 to 40 liters per minutes. Water level are within 25 meters to the surface. Success rates of 65 - 70% are achieved, following the geophysical survey.

3.2 Discussion of Results.

Bauchi Metropolis is located in the high relief features and the run off is relatively high, while the infiltration rate is low. Figure 2 shows the annual rainfall pattern, which reveals that the maximum amount of rainfall occurs in the months of July and August, as a result of the gradual buildup of rain which essentially begins from March with trace of the rain, the soil profile acquires its maximum moisture by August, at this point, most of the shallow wells have their highest water levels. This contributes significantly to the ground water storage through infiltration and percolation process, the resource gains access through the basement fractures. The ground water demand at these periods are also reduced due to the high availability of surface water for all of human and plant need.The geological features, in addition to the adverse climatic condition of the area, control the ground storage. The area is underlain by the crystalline basement complex, where the occurrence of ground water is due largely to the development of secondary porosity and permeability by weathering or fracturing of the parent rocks (Schiffler 2000).

The temperature variation in the study area is characterized by minimum and maximum temperatures as shown in fig. 3 and 4 respectively. The least temperature of 29.4^0 c was recorded in the month of August. This corresponds with the period of highest rainfall where the ground got saturated, resulting into the cooling of the entire study area.

It can be noticed that from fig.5, the relative humidity of this area is equally highest within the months of July and August this is owing to the high moisture level both on land and theatmospheric water vapour

Fig.6shows the chart of drawdown against time of the RafinZurfi area where the pumping test at the initial stage of zero-time depth is 5.1m, while pumping began after first 10 minutes of pumping, the drawdown is 36.4m, then the borehole stabilized at 120minute of starting, in which the rate of water extracted equal to the rate of water flow in to the borehole, that is the dynamic water level (DWL) of the Well. The total depth of the Well is 50m, which indicate that the borehole has the water column of 7.5m after stabilizing.

The result obtained in fig.7,show the residual drawdown against time in RafinZurfiwhere the water is rising after pumping stopped at the first 10 minutes, the water rouse from 12.5m at zero minutes to a stage of stability at the depth of 7.3m with the recovery period of 25minute. while at 6.0m, the time was at 30 minutes of stopping the pump. The water is about to become static with residual drawdown of 0.9m.

Thedrawdown of the borehole behind Nitel is also shown in fig. 8, in which at the first 10 minutes of the pumping test, the borehole stabilized at the depth of 8.6m, while the dynamic water level (DWL) attained, which the residual draw down in figure 9, shows the water column of the borehole is 36.4m, this indicate that the borehole has adequate water column and the aquifer is very productive.

From fig. 10, the borehole located opposite the hospital emergency unit shows significant change in the drawdown in the first 10 minutes.Water level rouse to 5.3m from the 8.60m, while at first 20 minutes the rise is at 3.0m thus, attaining static level up to 30 minutes at which the water is very sluggish at 2.7m. On the other handthe residual drawdown shown in fig. 11reveals the extraction of water at the first 10 minutes is at the 10.3m, while the borehole came to stabilized at the first 30 minutes at the depth of 11.1m with the total depth of the borehole at 44.9m, the water column is 33.8m which indicate the borehole has enough and very productive aquifer.

The borehole located opposite Niima hospital shown in fig.12 revealed how the water rouse after pumping stopped in relation to time taken. The pumping stopped at drawdown of 11.1m but after the first 18 minutes becomes static at 4.2m, this show that the aquifer is fast recharging since the well is

about to become static within the first 25 minutes of stopping the pumping operation as shown in fig.13

Fig. 14 show the drawdown of the borehole near Shafafilling station at Jos road. The extraction of water at the first 10 minutes is at 9.0m, while at 20 minutes the borehole attained the dynamic water level at 9.5m.The total depth of the borehole is 28m, the water column is 18.5m, which show that the borehole has enough water to supply at the flow rate of 1.0 l/sin 3 hours of pumping.

From fig.15, the residual drawdown after stopping rouse at a very high speed with the drawdown from 9.5m to 3.5m at first 10 minutes, then 3.4m at first 18 minutes of stopping pumping. This indicate that the borehole has a high recharge, since the flow rate in the last 2 hours is 1.0 l/s, which shows a quick recharge from the aquifer.

The result obtained and interpreted reveal that all the Vertical Electrical Sounding (VES) have potential for ground water exploration due to low resistivity and high conductivity in those VES points. From the study area, the resistivity value ranges from 24.8 to 847.1 ohms-meter, the nature and lithology units are. Topsoil, highly weathered basement, weathered basement and fresh basement.

Also, the basement complex formation in the study area, is mainly weathered and in some portion, the weathering is not too deep, e.g., quartzite is in the weathered zones, fractured and joints are also found on the basement in the study area.

The average depth to water table in the crystalline basement in the study area is 4.2m and the average yield of 0.87 l/sec. (52.2 l/min.).

The average transmissibility in the study area was found to be $7.90X10^{-3}$ m^2/sec.which happens to be of negligible potential according to Kumar, (2003), and their permeability is about moderate to low, sometimes very small.

Finally, the storability in the basement complex ranges from $1.002x10^{-3}$ to $1.008x10^{-3}$**.** This valueis reasonable, and when abstracted it can meet the demand of the populace.

IV. Conclusion

Groundwater recharge in Bauchi Metropolis solidly depend on the rainfall and to some degree inflowing seepage and leakage from other sources. In some areas, it is influenced by proximity nearest to rivers. It was observed that Aquifers with good recharge may have minor drawdown with high recovery rates. Also, the average depth to water table in the crystalline basement in the study area is 4.2m and the average yield of 0.87 L/sec. (52.2 L/min.), this was determined from the results obtained in the area. The average transmissibility in the study area is $7.90X10^{-3}$ m^2/secwhich happen to be of negligible potential, and their permeability is about moderate to low, sometimes very small. The storability in the basement complex is at reasonable ranges from $1.002x10^{-3}$ to $1.008x10^{-3}$. The average aquifer thickness in the area study area was approximately 2m, and the boreholes depth range from 30 to 50m. Generally, Bauchi Metropolis is located on the crystalline Basement which comprised moderately and poorly decomposed gneisses, granite, fractured pegmatite and quartz, based on that, the geophysical survey result shows Bauchi metropolis has potential ground water that can sustain the metropolis using Boreholes, to complement the supply of water from the main treatment plant by the State Water Board.

REFERENCES

[1.] Edet, A. E, Okereke, C. S., (1997). Assessment of Hydro-geological Conditions in Basement Aquifers of the Precambrian Oban Massif, Southeastern Nigeria. Journal of Applied Geophysics 36; 188:204.
[2.] Mohammed, I. N. Aboh, H. O., &Emenike, E. A. (2007). A Regional Geoelectric Investigation for Groundwater Exploration in Minna area, North West Nigeria. Science World Journal, 2(4), 15-19.
[3.] Adesanya, A.O, Okolie, E.C., (1991). Determination of Groundwater Potential in Obiaruku and Environs Using Surface Geoelectric Sounding. Environmentalist, 2006, 26, 301–308
[4.] Oyawoye M. O. (1963).The basement Complex of Nigeria. A. J. Whiteman(ed), African Geology, University Press, Ibadan, Nigeria.
[5.] Leduc, C., (2009). Land Clearing, Climate Variability, and Water Resources Increase in Semiarid Southwest Niger: AReview. Wat. Resource 45(7), W00A16.
[6.] Chilton, P.J., Moench, M., Cardy, F., Schiffler, M; (2000). Groundwater in Rural Development. World Bank Technical Paper 463, The World Bank, Washington DC.
[7.] Shemang EM, Jiba K.T, (2005). Hydro geo-electrical Study in the North-Eastern part of Adamawa State, Nigeria. Journal of Environmental Hydrology., p.14.Offix (1990) Resistivity Interpretation Program. Interpex limited, Golden Colorado, USA.
[8.] Kumar D, Ahmed S, (2003). Seasonal Behaviour of Spatial Variability of Groundwater Level in a Granitic Aquifer in Monsoon Climate. Current Sci., 84(2), 188-196.