

# ዘዴ

የኢትዮጵያ ሙያዊና ቴክኒካዊ ስራዎች ምዝገባ

ISSN: 0514-6216

Indexed on  
AJOL

# zede

**Journal of Ethiopian Engineers and Architects**

**40**

ሐምሌ  
July

፳፻፲፬  
2022

Annual Publication of the Addis Ababa Institute of Technology  
Addis Ababa University

ዘዴ

የኢትዮጵያ መሳሪያዎችና አርኪቴክቶች መጽሔት

**Zede**

Journal of Ethiopian Engineers and Architects

የተቋቋመው ፲፱፻፺፯

Established 1963

ሐምሌ ፳፻፲፬

July 2022

**40**

ፖ.ሣ.ቁ አዲስ አበባ

P.O.Box 385 Addis Ababa

**Editor-in-Chief**

*Abebe Dinku*

**Asso. Editor**

*Mohammed Abdo*

**Editorial Board Members**

*Adil Zekaria*

*Edessa Dribssa*

*Getachew Bekele*

*Heyaw Terefe*

**Publisher**

Addis Ababa University

Addis Ababa Institute of

Technology

P.O.Box 385

Addis Ababa

Ethiopia

**Postal Address**

Addis Ababa University, AAiT

P.O. Box 385

Addis Ababa

Ethiopia

**Website:** [www.aait.edu.et/zede](http://www.aait.edu.et/zede)

**Email:** [zede@aait.edu.et](mailto:zede@aait.edu.et)

**CONTENTS**

**Pages**

1. Application of Shallow Energy Piles for the Purpose of Heating and Cooling of Building  
**By: Henok Fikre** 1
2. Fat or Elastic: an Inquiry into Classification of Black Clay Soils of Addis Ababa around Tulu Dimtu Area  
**By : Tewodros Gemechu** 11
3. Bacterial Contamination of Schools Drinking Water in Addis Ababa, Ethiopia  
**By: Dawit Debebe Zerihun Getaneh, and Fiseha Behulu** 23
4. Analysis of Track Geometry Index Measurement Methods  
**By: Daniel H/Michael , Elias Kassa and Getu Segni** 33
5. Analytical Study on Seismic Performance of Partially Prestressed Concrete Joints  
**By: Hilina Assega and Adil Zekaria** 45
6. Mechanical and Durability Properties of Concrete with Granite Powder as Partial Replacement of Cement  
**By: Asnake Kefelegn , Binaya Patnaik and Issayas Kebede** 59
7. Distance Aware Transmit Antenna Selection for Massive Mimo Systems  
**By: Shenko Chura, Yalemzewd Negash and Yihenew Wondie** 73
8. Addis Ababa Light Rail Transit System Energy Flow Analysis  
**By: Asegid Belay and Getachew Biru** 87
9. Detection and Restoration of Click Degraded Audio Based on High Order Sparse Linear Prediction  
**By: Bisrat Derebssa ,Eneyew Adugna ,Koean Eneman and Too Van** 99

THE EDITORIAL BOARD IS NOT RESPONSIBLE FOR VIEWS EXPRESSED BY INDIVIDUAL AUTHORS.

## Guide to Authors

ZEDE is a scientific journal on engineering science and application, produced under the auspices of the Addis Ababa Institute of Technology, Addis Ababa University. The main objective of the journal is to publish research articles, findings and discussions on engineering sciences, technology and architecture thereby assisting in the dissemination of engineering knowledge and methodologies in solving engineering problems. Technical Notes of significant contribution may be considered for publication.

Original papers for publication in the journal should be submitted in triplicate to the Editor-in-Chief, P.O. Box 385, Addis Ababa, Ethiopia. All articles submitted for publication in the journal should comply with the following requirements:

**1. Title of Paper:** The title of the paper should be phrased to include only key words and must have a length of not exceeding 80 characters including spaces.

**2. Format of Manuscript:** The manuscript should be (double-spaced single column draft and single spaced double column final) in A-4 sized paper with MS word 2007 or later version. Margins of 25 mm should be used on all sides of the paper.

**3. Length of Article:** The length of the article should not exceed word equivalent of 6000 words, or 20 pages, double spaced using font size 12 typed in Times New Roman.

**4. Author's Affiliation:** The author's full name, institutional affiliation and rank, if applicable, must appear on the paper.

**5. Abstract:** All articles submitted must include an abstract of length not exceeding 200 words in *italics*.

**6. Keywords:** All articles submitted must include Keywords not exceeding 6 in number.

**7. Style of Writing:** It is recommended that third person pronoun/s be used when referring to author/s.

**8. Illustrations:** Figures should be drawn in black, at a size with a 50% reduction to fit in 160 mm width of journal. Photographs should be submitted as glossy prints. Explanations and descriptions must be placed in the text and not within figures. All figures must include numbered captions. See example:

Figure 1 Typical creep strain versus time curve

**9. Tables:** Tables must be numbered in the same order as cited in the text. Explanations of tables must appear in the text.

**10. Equations:** Equations numbers should be right-justified. See example:

$$u(x, y) = -y\theta(x) \quad (1)$$

**11. References:** References in the body of the Article should be cited at the end of the paper by placing a reference number in square brackets and should be arranged sequentially as they appear in the text. Ethiopian names may be given in direct order, i.e. given name followed by father's name. All main words in titles (papers, books, reports) should be initialized by capital letters. Items in citations should be separated by commas. Page numbers should be included whenever applicable

### Examples:

#### 1. References to Journal Articles and Proceedings

Spillers, W.R. and Lefeoehilos, E.,  
"Geometric Optimization Using Simple  
Code Representation", Journal of the  
Structural Division, ASCE, vol. 106, no.  
ST5, 1980, pp. 959-971.

#### 2. References to Books and Reports

Korsch, H.L. and Jodl, H. -J.,  
"Chaos: A Program Collection for the  
PC", Springer-Verlag, 1994.

**12. Units:** SI units must be used.

**13. Conclusions:** A set of conclusions must be included at the end of the paper.

#### 14. Submission of Paper:

Any paper submitted for publication in ZEDE must not have been published previously, or submitted for publication elsewhere; and if accepted for publication by ZEDE, the author/s shall transfer the copy right to ZEDE.

# APPLICATION OF SHALLOW ENERGY PILES FOR THE PURPOSE OF HEATING AND COOLING OF BUILDINGS

**Henok Fikre**

School of Civil and Environmental Engineering, Addis Ababa Institute of Technology,  
Addis Ababa University, Addis Ababa, Ethiopia  
E-mail: [henok.fikre@aaait.edu.et](mailto:henok.fikre@aaait.edu.et)

## ABSTRACT

*Shallow energy piles are these days internationally being implemented for serving as heat exchanging media between the ground and the building in addition to supporting structural loads. This research aims at providing the basic concepts for the application of energy piles for a portion of the total demands of high-rise buildings in Ethiopia, specifically heating and cooling demands. Preliminary estimation of the number of energy piles to cover a certain portion of the total heating and cooling demands of a selected building in Addis Ababa showed the potential for its application in the future. Investigation of the behaviour of energy piles due to the action of cycling heating and cooling loads by using finite element software TOCHNOG were also performed by employing the appropriate hypo-plasticity constitutive model for the soil layers. The measured values of the thermo-mechanically loaded pile were reasonably simulated by the finite element model. The coupled thermo-mechanical cyclic thermal and mechanical loads were found to reduce the settlements of the foundation significantly, which can be considered as an additional advantage of this foundation technique for high rise buildings.*

**Key words:** energy piles, finite element analysis, HVAC, hypo plasticity.

## INTRODUCTION

In this world of increasing development activities due to urbanization, the energy demand is growing rapidly. Since most of the traditional sources of energy are associated with environment pollution, it is desirable to seek for more modern and renewable energy sources. Geothermal energy is among such sources, which is capable of minimizing the carbon-dioxide (CO<sub>2</sub>) emission and promotes compliance with international environment obligations such as the Kyoto and Toronto targets [1, 2].

There are two common types of applications of geothermal energy, namely deep geothermal and near surface geothermal technologies. While deep geothermal technologies exploit deep reservoirs with temperatures higher than 100°C, near surface geothermal technologies exploit heat energy stored at shallow depth, i.e. commonly not deeper than approximately 200 m. Down to a depth of approximately 5 m, thermal energy absorbed from solar radiation energy plays a significant role. Since the groundwater and soil particles have high heat storage capacities, it is nowadays common practice to use the shallow sources for a certain portion of the energy demand of buildings through structural elements such as energy piles and retaining walls [3]. The good thermal storage and conductivity of concrete makes these structures ideal heat exchange media.



High density polyethylene pipes, in which a heat carrier fluid circulates, installed within the concrete structures extract the geothermal energy from the ground [4].

According to [4, 5], the ground temperature tends to be constant below approximately 10 - 15 m with the magnitude depending on the location (10 – 15°C in Europe and 20 - 25°C in Africa). These temperature ranges are sufficient to allow heating and cooling of

buildings in winter and summer respectively. Thermal structures like energy piles are frequently incorporated in to buildings in Austria, Germany and Switzerland [1, 6]. The technology is presently attaining more practical application all over the world [8 - 10]. Some of the successful applications of energy piles for covering the heating, ventilation and air conditioning system (HVAC) demands of buildings have been summarized in Table 1.

Table 1. Worldwide energy pile applications for covering the HVAC demands of buildings

Building	Location	Total no. of piles	Energy piles	Dimension	Serving
Lainzer tunnel	Vienna	59	1/3 of total	1.2 m diam. 17.1 m length	Heating/cooling (214 MWh)
SFIT	Lausanne	440	300	0.9-1.5 m diam. 30 m length	85% of HVAC demand
HochVier	Frankfurt	302	212	1.86m diam. 27 m length	HVAC
Keble building	Oxford	61	all	0.45 m diam. 12 m length	full HVAC demand

Application of the energy piles at Swiss Federal Institute of Technology in Lausanne and Dock Midfield Airport in Switzerland showed that the additional cost of implementing energy piles has already been compensated with the energy savings only within eight years [11]. Based on these and other experiences, a preliminary analysis has been performed regarding the application of shallow energy piles in Addis Ababa based on the demand analysis of a selected high-rise building, to cover a certain portion of its heating and cooling demands.

Understanding the effects of temperature variations on the mechanical behaviour of the foundation and the ground is a key factor for an optimized application of energy piles in any parts of the world. To that effect, many in-situ tests have been performed on different ground conditions to reveal the behaviour of energy piles [12, 13]. Even if

there have been research activities to model the behaviour of energy piles using numerical tools, there are still gaps to acquire material models which represent the cyclic loading scenario with close proximity to real behaviour depicted during in-situ measurements. This research further addresses the behaviour of a thermo-mechanically loaded pile in layered soils, due to cyclic thermal loading using finite element method by applying appropriate constitutive model for the soil by validating the experimental results of Laloui et al. [11].

### **Energy Demands in Ethiopia and introduction of energy piles for HVAC demands**

The Ethiopian energy sector faces the dual challenges of limited access to modern energy and heavy reliance on traditional biomass energy sources to meet the ever-

growing demands, which is associated with environment pollution [14]. In recent years, despite the fact that the country has been in continuous economic growth, there is a challenge to get the required energy supply to sustain this growth into the future.

According to Mondal et al. [14], more than 80% of the energy in Ethiopia remains to be

consumed by rural and urban households until 2030 as shown in Fig. 1, dominantly meant for cooking and heating purposes. However this will be associated with environment pollution due to the tremendous CO<sub>2</sub> emission. It is thus important to look for sustainable renewable energy sources targeted to cover a significant portion of this demand.

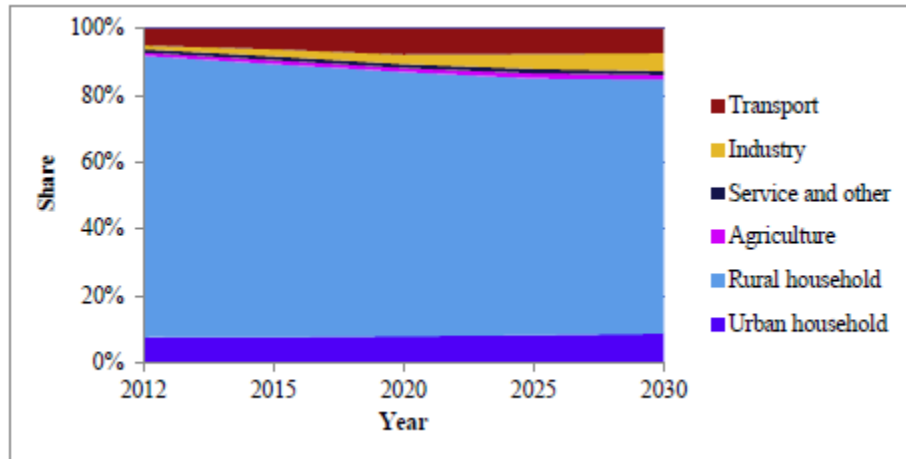


Fig. 1 Sector-wise percentage share of energy consumption in Ethiopia [14]

The percentage shares of urban household, transport and industry sectors show gradual increase with time while that of rural household decreases, due to development activities and increased urbanization. These sectors have heating and cooling demands in addition to the basic energy needs.

As experienced in the other parts of the world, shallow geothermal energy can be considered as a viable renewable energy source for industry as well as household purpose, which can be associated with the foundation elements. If piles are accompanied by energy accessories from the ground, the energy production doesn't depend on the building area in contrary to solar energy, rather on the pile contact area with the ground, which is entirely dependent on the pile geometry.

Even if the energy extracted from the pile foundation shall be calculated based on detailed three - dimensional analyses, Brandl [15] provided design guidelines for general feasibility studies which can show the potential of using energy piles. Accordingly, the energy volume that can be extracted from thermo active energy piles can be estimated as a function of the pile diameter,  $D$ , as follows:

- pile foundations, with piles  $D = 0.3 - 0.5$  m: 40 - 60 W per meter run,
- pile foundations, with piles  $D > 0.6$  m: 35 W per m<sup>2</sup> earth-contact area

In the design and analysis of energy foundation, one of the first important tasks is analysing the energy demands of the building for HVAC demands. This value is considered when deciding the dimension of the foundation. Spacing, diameter & length

as well as the number of piles with assigned dimensions will be evaluated if the energy extracted fulfils the cooling and heating requirement of the building.

According to [16], the heating and cooling demands of residential and commercial buildings in Africa accounts for about 20 % and 52 % of their total energy demands, respectively. Although detailed analyses need to be performed, the Author of this research estimates the HVAC demand of buildings in Addis Ababa to be 20 – 30 % of the total energy demands. This research is aimed at demonstrating the possibility of covering this amount of energy demands by the use of energy piles. The already constructed 4B+G+32 United Bank Head Quarter building was chosen for the preliminary demand and supply analysis of

energy from shallow geothermal sources in Addis Ababa. The structure was built on 3,338 m<sup>2</sup> area of land and the main tower of the building was supported by 282 cast in-situ reinforced concrete piles of 28 m length and 0.8 m diameter spaced at 2.4 m. The contact area of a single pile with the soil is thus found to be 70.37 m<sup>2</sup>, which is used for the estimation of the required number of piles. The total energy demands of the building was assessed to be 2,698 KVA; among which, 761.44 KVA or 609.15 kW, is required for heating and cooling (HVAC) of the building, which accounts for 28.22 % of the total demands. The number of energy pile required for fulfilling the cooling and heating demands of the building are determined according to the aforementioned pre-design method of Brandl [15] as illustrated in Table 2.

Table 2: Comparison of the amount of energy extracted from different number of energy piles

<b>No. of Energy Piles</b>	<b>% Energy Pile/ Total No. Piles</b>	<b>Total Contact Area(m<sup>2</sup>)</b>	<b>Amount of energy extracted (kW)</b>	<b>Energy required for HVAC system(kW)</b>	<b>%Extracted/Required</b>
<b>50</b>	17.73	3518.58	123.15	609.15	20.22
<b>75</b>	26.60	5277.87	184.73	609.15	30.33
<b>150</b>	53.19	10555.74	369.45	609.15	60.65
<b>200</b>	70.92	14074.32	492.60	609.15	80.87
<b>249</b>	88.30	17522.53	613.29	609.15	100.68

It can be observed that the total HVAC demand of the building can be covered by using 249 energy piles, i.e., 88.3 % of the total number of structural piles. It is also possible to produce a portion of the total HVAC demand of the building by reducing the piles. Consequently, half and a quarter number of the total number of piles could produce about 60 % and 30 % of the total HVAC demand, respectively.

### **1 . Predicting settlements of energy piles due to coupled cyclic thermo-mechanical actions**

Understanding the behaviour of energy piles due to the combined actions of mechanical and cyclic thermal loads in layered soils is the key for the successful application of the technology. Numerical methods are among the approved methods of analysis for such type of complex structures in the European code [17]. Owing to its layered formation the in-situ measurements of [11], performed

on one of the piles of a four-story building at Ecole Polytechnique Federal de Lausanne (EPFL), has been considered for further analysis.

The building with a ground area of 100 m x 30 m is founded on 97 bored piles with a pile length of  $L_p = 25.8$  m and a diameter of  $d_p = 0.88$  m. The groundwater level is located close to the ground surface. The mechanical load acting on the test pile corresponds to the dead weight of the building under construction. The thermal load was induced by a heating device controlling the temperature of the water used as heat carrier fluid in the PE tubes. The two types of loads were applied separately and

alternately in order to identify thermal and mechanical effects. As shown in Fig. 1, the thermo mechanical loading was depicted in seven cycles, excluding Test0 which represents the measurements made during the casting of the pile. During the first phase (Test1), the temperature is increased to a maximum value of 21.8 °C beyond which unloading follows to the minimum value. The mechanical load of  $Q = 1300$  kN was applied linearly beginning from the end of the first step to the end of the 6<sup>th</sup> cycle. At the end of the construction of each story, a thermal loading cycle was applied to a maximum of  $\Delta T = 15$  °C as schematized in Fig. 1.

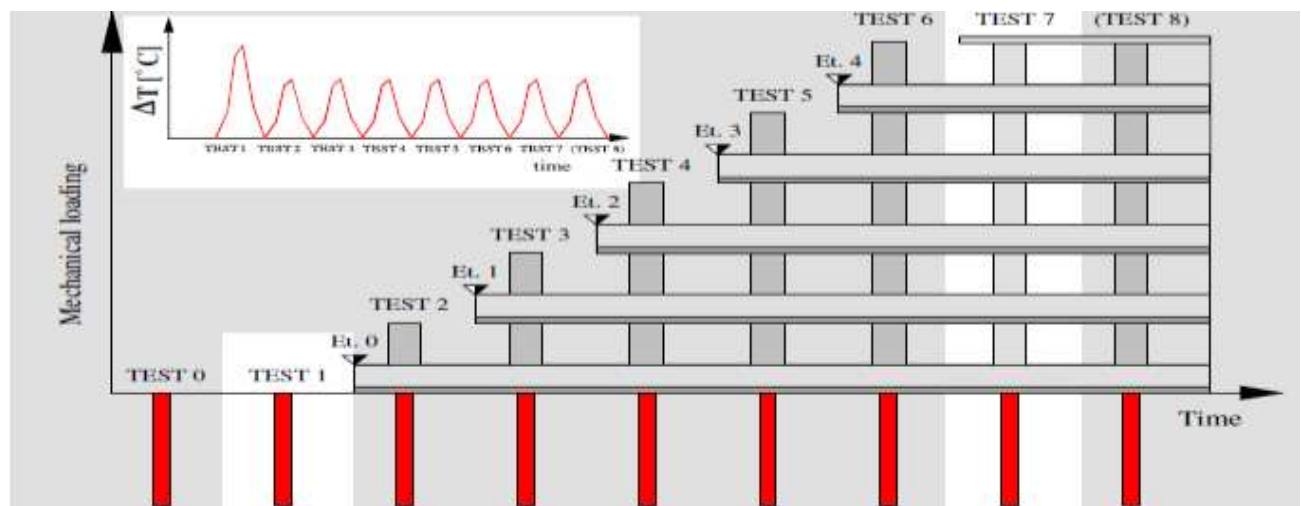


Fig. 2 Thermo-mechanical loading history (after [11]).

The results of the measurements have been compared with those of an axis-symmetric finite element model employed in TOCHNOG software [18]. The finite element mesh presented in Fig. 3 comprises approximately of 4900 quadrilateral elements, together with appropriate boundary conditions in the model for all loading conditions.

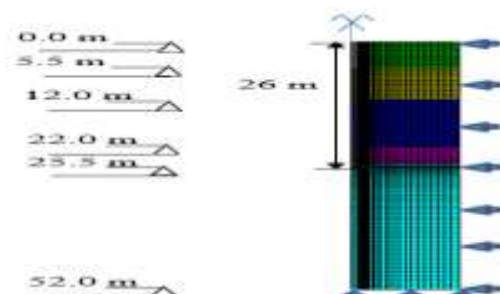


Fig. 3 Finite element mesh

Based on the results of a preliminary study comparing various constitutive models, the hypo

plastic model by Masin [19] has been adopted for the predominantly cohesive soil layers. Masin's model incorporates the concept of intergranular strains first proposed by Niemunis and Herle [20], which considers the small strain stiffness allowing simulation of some effects of cyclic soil behaviour. The parameters for the

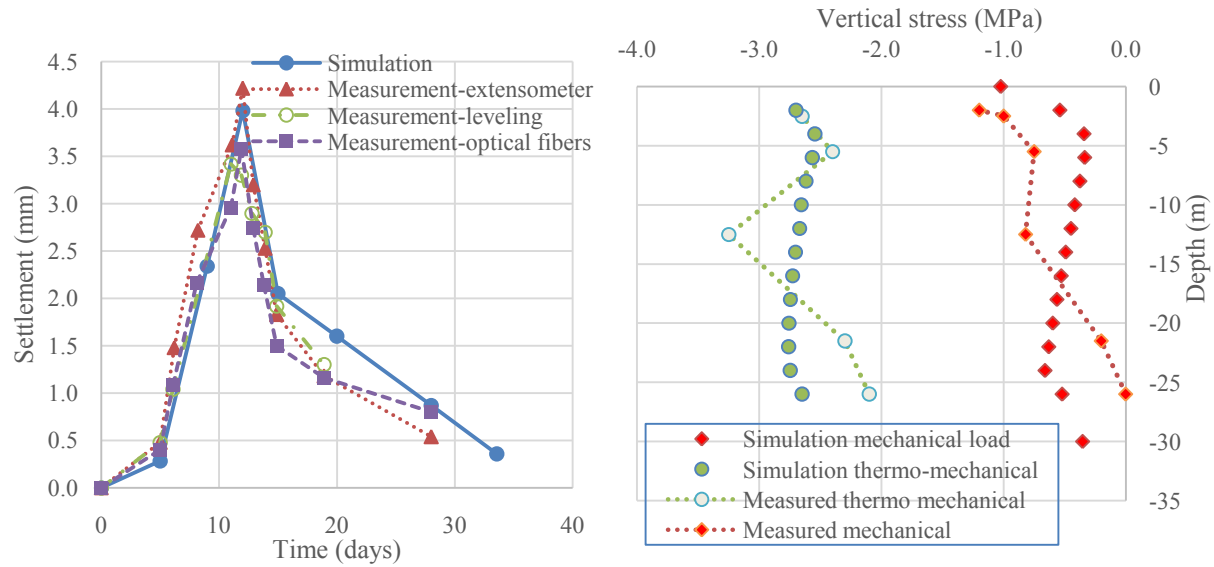
selected constitutive model have been derived from the reports of [14, 21, 22] and the concrete pile is assigned linear-elastic parameters with Young's modulus,  $E$  of 29.2 MPa and Poisson's ratio,  $\nu$  of 0.25. The basic soil parameters used in the analyses are summarized in Table 3.

Table 3 .Basic material parameters for the soil layers

Material parameters for the respective layers	A1 Alluvial soil	A2 Alluvial soil	B Sandy gravelly moraine	C Bottom moraine	D Molasse
Poisson's ratio $\nu$ [-]	0.2	0.2	0.4	0.4	0.3
Elastic modulus $E$ [MPa]	190	190	84	90	3000
Friction angle $\phi'$ [°]	30	27	25	27	35
Cohesion $c$ [kPa]	5	3	6	20	2000
Thermal conductivity $\lambda$ [W/m °C]	3.38	3.38	4.17	4.17	2.38
Specific heat capacity $C_s$ [Joules/m <sup>3</sup> °C]	2463.7	2463.7	2434.2	2438.6	2359.2
Thermal expansion coefficient of solid state, $\beta_s$ [per °C]	$3.3 \times 10^{-6}$	$3.3 \times 10^{-6}$	$3.3 \times 10^{-6}$	$3.3 \times 10^{-6}$	$3.3 \times 10^{-5}$
Thermal expansion coefficient of liquid state $\beta_w$ [per °C]	$2.0 \times 10^{-4}$	$2.0 \times 10^{-4}$	$2.0 \times 10^{-4}$	$2.0 \times 10^{-4}$	$2.0 \times 10^{-4}$
Ground flow capacity $C_w$ [Joules/m <sup>3</sup> °C]	4186	4186	4186	4186	4186
Permeability $k$ [m/s]	$2 \times 10^{-6}$	$7 \times 10^{-7}$	$1 \times 10^{-6}$	$1 \times 10^{-6}$	-

The thermo-mechanical loading scheme of the in-situ test has been analyzed by considering the step-by-step procedure of the practical condition as presented in Fig. 1. Two sets of measured data (at the 1<sup>st</sup> and

7<sup>th</sup> cycles) have been used for validating the numerical model. Fig. 4 a) and b) present the comparative results of the measured and computed values for the displacements of the pile head in the first cycle and the vertical stress at pile shaft in the 7<sup>th</sup> cycle respectively.



a) Settlement versus time(1<sup>st</sup> cycle)

b) vertical stress with depth (7<sup>th</sup> cycle)

Fig. 4 Comparison of measured and numerically computed values

Since both simulations of the mechanical load and the thermo - mechanical load are in close agreement with the measured values, further analyses regarding the effects of cyclic loading on the settlements of an energy pile have been performed by employing the proposed constitutive model

for the soil layers. The assumptions and analyses phases used for the validation purpose have been adopted for the analyses. The influence of cyclic heating and cooling on the pile head displacement is depicted in Fig. 5.

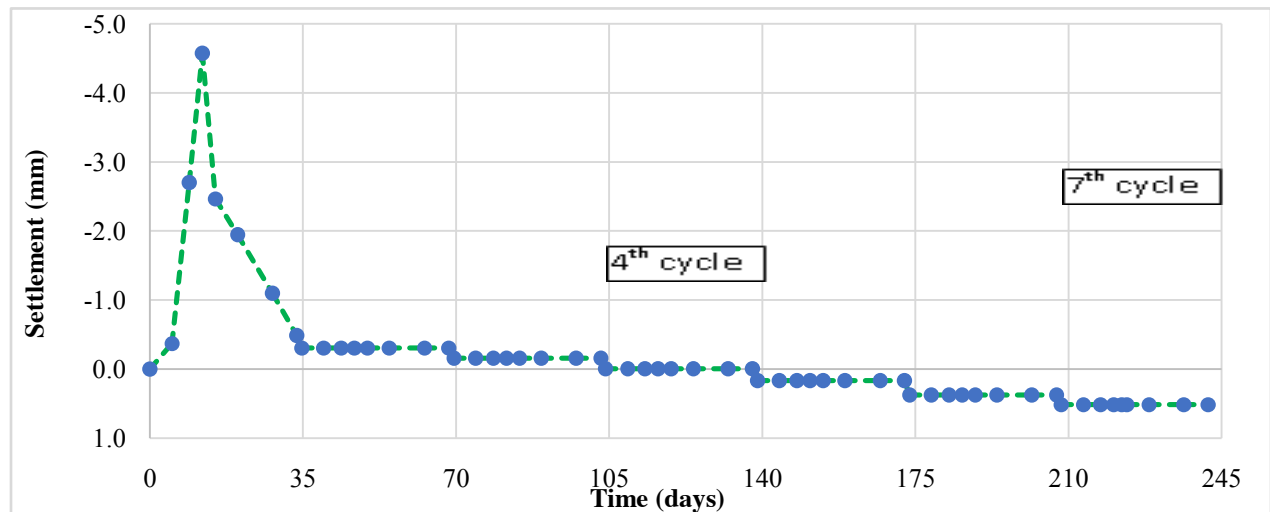


Fig.5 Effect of cyclic thermo mechanical loading on pile head settlement



In the first cycle, where temperature alone is varied without the mechanical load, elongation of the raft is found to be proportional to the applied thermal load, the maximum value being recorded at the maximum temperature of 21.8 °C. The elongated pile doesn't return to the original position while unloading the thermal load in this cycle, except in the fourth cycle, after about 40 % of the mechanical load is applied together with the thermal load cycles. The dominance of the mechanical load is evident in all the cycles following the first (thermal loading cycle), where the variation of the temperature has not been reflected on the settlement. This is an indicator of the advantage of having energy piles to reduce the settlement of high-rise buildings.

### CONCLUSIONS

Due to the ever increasing demand of energy, this research has focused on introducing the idea of shallow energy foundations for the demands of heating and cooling purposes of high-rise buildings in Ethiopia. Based on energy demand analysis of some of the buildings in Addis Ababa, a preliminary estimation of the number of energy piles for covering the heating and cooling demands of the already constructed United Bank office building has been performed. It was found out that only 88 % of the total number of structural piles could sufficiently satisfy the HVAC demand of the same building.

For the purpose of understanding of the thermo-mechanical behaviour of energy piles for introducing of energy piles in our country, numerical analysis of a practical application of energy piles in layered soils has also been incorporated in the research. After preliminary analyses using different constitutive models, the hypo plasticity model of Masin has been found to simulate the cycling loading reasonably with the computed values showing very good

agreement with the measured ones. The induced settlements of the foundation were also found to be reduced significantly due to the coupled thermo-mechanical loading.

### REFERENCES

- [1] De Moel, M., Bach, P., Bouazza, A., & Sun, J. (2010). Technological advances and applications of geothermal energy pile foundations and their feasibility in Australia. *Renewable and Sustainable Energy Reviews*, 14(9), 2683 – 2696.
- [2] Freedman, M., Freedman, O., Stagliano, A.J. (2015). Assessing CO2 Emissions Reduction: Progress toward the Kyoto Protocol Goals in the European Union. *International Journal of Business and Social Research*, 5(11): 75-86.
- [3] McCartney, J.S. (2016). Parameters for load transfer analysis of energy piles in non-uniform plastic soils. *Int. J. Geomech*, 1532-3641.
- [4] Brandl H. (2006). Energy foundations and other thermo-active ground structures. *Géotechnique*, 56(2):81–122.
- [5] Singh, R.M., Bouazza, A., Wang, B. (2015). Near-field ground thermal response to heating of a geothermal energy pile: Observations from a field test. *Soils and Foundations*, Elsevier, pp. 55(6): 1412-1426.
- [6] Von der Hude, N., Sauerwein, M. (2007).Energiepfähle in der praktischen Anwendung. *Mitteilungen des Institutes und der Versuchsanstalt fuer Geotechnik der Technischen Universitaet Darmstadt*, Heft Nr. 76, 95-109.

- [7] Amatya B.L., Soga K., Bourne-Webb P.J., Amis T., Laloui L. (2012). Thermo-mechanical behaviour of energy piles, *Geotechnique*, 62, No. 6, 503-519.
- [8] Bourne-Webb, P.J. (2013). A framework for understanding pile behaviour. ICE Publishing, 170-177.
- [9] Knellwolf, C., Peron H., and Laloui, L. (2011). Geotechnical Analysis of Heat Exchanger Piles. *Jnl. of Geotechnical and Geo environmental Engineering* 137(10), pp. 890-902.
- [10] Ng, C. W. W., Shi, C., Gunawan, A., Laloui, L. (2014). Centrifuge modelling of energy piles subjected to heating and cooling cycles in clay. *Geotechnique Letters* 4, 310–316.
- [11] Laloui, L. (2006). Experimental and numerical investigation of the behaviour of a heat exchanger pile. *Int. Jnl. for Numerical and Analytical Methods in Geomechanics*, 30(8), pp. 763-781.
- [12] Wang C.L., Liu H.L., and Kong G.Q. (2016). Model tests of energy piles with and without a vertical load. *Environmental Geotechnics*, 3(4):203–213.
- [13] Yavari N, Tang A.M., Pereira J.M., Hassen G. (2014). Experimental study on the mechanical behaviour of a heat exchanger pile in physical model. *Acta Geotechnica* 9, 385-398.
- [14] Mondal, M.A.H., Bryan, E. Ringler, C., Mekonnen, D. (2018). "Ethiopian energy status and demand scenarios: Prospects to improve energy efficiency and mitigate GHG emissions". *Energy*, 161-172.
- [15] Brandl H. (2016). Geothermal Geotechnics for Urban Undergrounds. *Procedia Engineering* 165: 747 – 764.
- [16] Ürge-Vorsatz D., Cabeza L.F., Serrano S., Barreneche C., Petrichenko K. (2015). Heating and cooling energy trends and drivers in buildings. *Renewable and Sustainable Energy Reviews* 41, 85-98
- [17] CEN European Committee of Standardization, Eurocode 7: Geotechnical design - Part 1: General Rules (2004).
- [18] Tochnog, (2019). TOCHNOG Professional – Finite Element Analysis. Tochnog Professional Company, Nijmegen, The Netherlands.
- [19] Masin, D. (2014). Clay hypoplasticity model including stiffness anisotropy. *Geotechnique*, 64(3), pp. 232-238.
- [20] Niemunis, A., Herle, I. (1997). Hypoplastic model for cohesionless soils with elastic strain range. *Journal of Mechanics of Cohesive-Frictional Materials*, 4, Issue 2, pp. 279-299.
- [21] Di Donna A., Loria R., Laloui L. (2015). Numerical study of the response of a group of energy piles under different combinations of thermo-mechanical loads. *Computers and Geotechnics*, Vol. 72, No. 2016, 126–142.
- [22] Rotta Loria, A. F., Laoui, L. (2017). Group action effects caused by various operating energy piles. *Geotechnique*, vol 17: 213.



# FAT OR ELASTIC: AN INQUIRY INTO CLASSIFICATION OF BLACK CLAY SOILS OF ADDIS ABABA AROUND TULU DIMTU AREA

**Tewodros Gemechu**

<sup>1</sup>School of Civil and Environmental Engineering, Addis Ababa Institute of Technology (AAiT), Addis Ababa University, Ethiopia

\*Corresponding author: [tewodros.gemechu@aait.edu.et](mailto:tewodros.gemechu@aait.edu.et)

## ABSTRACT

*Addis Ababa Black clays are one of two dominant clay found in the city and known for its expansive nature. Using the unified soil classification system, such soil, traditionally, is classified by the group name fat clay and group symbol of CH. But some investigations on local soil have resulted in classifying the soil as elastic silt, MH.*

*In this research laboratory investigation is conducted to determine if in fact Addis Ababa black clays may end up being classed as elastic silts and if so, what may be the parameters that play a role in such a classification. For the study three samples were collected from Tulu-Dimtu, where previous investigations have resulted in the discrepancy. Simple classification tests were conducted on the samples. In addition, the effect of sample preparation, utilization of tap water, experience level of operator and variations among laboratories investigated.*

*It was found that Addis Ababa black clay soils may end up being classified as elastic silts but in general remain within the boundary region of the A-line on both sides. The experience level of the operator was found to have the most profound effect on index tests and classification.*

**Keywords** Black clay, liquid limit, plastic limit, USCS

## INTRODUCTION

Currently, two major groups of soil classification systems are available for general engineering use. They are those based on Arthur Casagrande's unified soil classification system (USCS) and those used for specific purposes such as the American Association of State Highway and Transport officials (AASHTO) system for classification of subgrade soils for highway construction purposes. Both systems use simple index properties such as grain-size distribution, liquid limit, and plasticity index of soil (Carter and Bentley 2016).

Even though identification and classification systems specific to expansive soils exist (Chen 1975), engineering classification system such as the unified soil classification system are initially conducted and based on these further testing is done to ascertain expansiveness. It is generally taken that soils classed under CL or CH by the USCS and A6 or A7 by AASHTO may be susceptible to expansibility (Nelson and Miller 1997).

Addis Ababa black clay soil is one of the two dominant clay found in the city, known for its expansive nature. Using the unified classification system, this soil is commonly classified under fat clay (CH) (Alemayehu Teferra 1992; A. Teferra and Yohannes 1986). This correlates to soil whose characteristics are heavily dependent on moisture content, have very low

permeability, may be susceptible to swelling and shrinkage and have high compressibility.

But some investigations conducted in Addis Ababa indicate, the soil is classified as an elastic silt (MH), which fall below the A-line at a liquid limit greater than 50 in the Casagrande plasticity chart. This correlates to silty soils of high plasticity which is a departure from the commonly accepted classification of expansive soils.

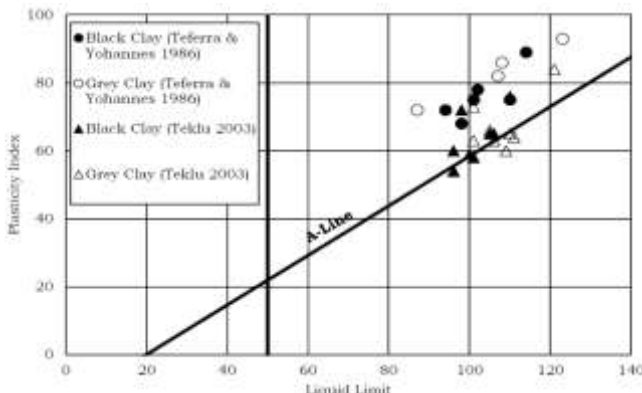


Figure 1 Plasticity chart plot of data from Tefera&Yohannes (1986) and Teklu (2003)

A survey of available literature shows that there is precedent for soils with high swelling and shrinkage characteristics, to be classified as elastic silts. In Israel, where shrinking soils are known to occur, plasticity data (LL and PI) plots below the A-line (Smith et al. 1985).

In a study on expansive soils in Sudan it has been found that such soil plot below the A-line in the elastic silt region (Al Haj and Standing 2015). In a study on Ethiopian soil, it has been reported that Ethiopian black clays plot near or below the A-line grouping them as elastic silts (Morin and Parry 1971). In Addis Ababa, a study conducted on expansive soils reports that plasticity data that plots below the A-line (Teklu 2003).

### Laboratory tests used for engineering classification of soil

The simple tests by which the various types of soil are identified and classified for geotechnical engineering use are called index or classification tests and the properties that are associated with them index properties (Terzaghi, Peck, and Mesri 1996; Das and Sobhan 2017).

For most of the common classification schemes both the consistency limits and the particle size distribution are required for classification of a soil. From the consistency limits the liquid and plastic limits are used.

The liquid limit, theoretically, is the transition point on the water content line from a plastic behavior to a liquid behavior. In practice, it is determined at a water content corresponding to arbitrary selected low shear strength (O'Kelly, Vardanega, and Haigh 2017). There exist two methods for the determination of the liquid limit, the Casagrande cup method and the fall cone method (Head 1992). The Casagrande cup method is standardized by the ASTM, AASHTO and BS (D18 Committee 2017; AASHTO 2013b; British Standards Institution 1990). The ASTM and AASHTO standards are equivalent while the BS standard defers from the rest in terms of the specification for the rubber base (O'Kelly, Vardanega, and Haigh 2017).

The Casagrande apparatus is based on Atterberg's initial method for the determination of the liquid limit which relied on the number of blows to close a groove in a soil bed to collapse when struck by hand. This is likened to the collapse of a slope which is related to the shear strength of the soil (Haigh, Vardanega, and Bolton 2013). Casagrande attempted to standardize the approach by specifying the liquid limit as the moisture content at which a groove cut in a soil bed and resting on a spun brass cup

closes at 25 blows for 13mm when the cup is impacted on a hardened rubber base from a height of 10mm at a rate of 0.5 blows/sec. The groove is cut using a standard grooving tool (Haigh, Vardanega, and Bolton 2013; Head 1992).

The fall cone apparatus also relies on the shear strength definition of the liquid limit but here penetration resistance is used as a measure. The liquid limit is defined as the moisture content at which a cone of mass 80g with an apex angle of 300 penetrates a soil specimen 20mm (British Standards Institution 1990; Head 1992). This method is standardized in the British standard and is the definitive method in the British standard (Head 1992). The primary advantage of the fall cone method is the reduction in variability. (O'Kelly, Vardanega, and Haigh 2017).

The plastic limit is the lower boundary moisture content for plastic behavior. A. Casagrande proposed rolling method which involves rolling a soil thread to a diameter of 3.2mm and observing for cracks (Haigh, Vardanega, and Bolton 2013). This method is standardized in ASTM, AASHTO and BS (D18 Committee 2017; AASHTO 2013b; British Standards Institution 1990). The method has multiple drawbacks which include its heavy reliance on operator judgment, variable rolling pressure and difficulty in assessing brittle cracking (Barnes 2009; Haigh, Vardanega, and Bolton 2013). Due to such drawbacks other methods have been proposed including the fall cone method where a strength definition of 100 times the shear strength as the liquid limit is considered (Sivakumar et al. 2009).

It has been shown that such an approach is in contradiction of the plastic-brittle transition definition of the plastic limit (Haigh, Vardanega, and Bolton 2013).

The plastic limit so determined is therefore not consistent with that determined from rolling and is designated as PL100 (O'Kelly, Vardanega, and Haigh 2017). Another method proposed involves utilization of a mechanical roller (Barnes 2009).

The liquid and plastic limit is influenced by sample preparation techniques, chemistry of water used, and soil fraction tested (O'Kelly, Vardanega, and Haigh 2017).

There exist standardized procedures for sample preparation (D18 Committee 2011a, 2011). There generally two, the dry method and wet method. The dry method is the preferred method for granular soil while, the wet method is recommended for fine grained soils especially those whose characteristics are changed by oven drying. The wet method discussed in ASTM D2217 involves the two methods one by air drying the other by washing.

### **Alternative Classification scheme**

There has been alternative soil classification schemes proposed for fine grained soil, especially relating to the Casagrande plasticity chart (E. Polidori 2003; Ennio Polidori 2015; Moreno-Maroto and Alonso-Azcárate 2018).

E. Polidori (2003) proposed a classification scheme that redefines clay and silt based on proportion of the clay fraction, fraction finer than  $2\mu\text{m}$ , computed from the portion of specimen finer than  $425\mu\text{m}$ , the fraction used for Atterberg limit testing.

According to this definition clay is a soil containing clay fraction greater than or equal to 50% while silt having clay fraction less than 50%. Using this definition and Atterberg limit tests of mixtures of montmorillonite and kaolinite with sand, the author proposed a new plasticity chart.



The Polidori plasticity chart, shown in Figure 2, consists of the C-line, 0.5C-line and the U-line. The C-line is plotted by connecting the plots of liquid limit versus the plastic index for the 100% clay fraction montmorillonite and kaolinite data tested. The 0.5C-line is plotted by connecting the test data for 50% clay fraction. The U-line represents the upper limit of expected behavior for natural soils. It is determined from data obtained for specimen with larger than 50% sand (E. Polidori 2003).

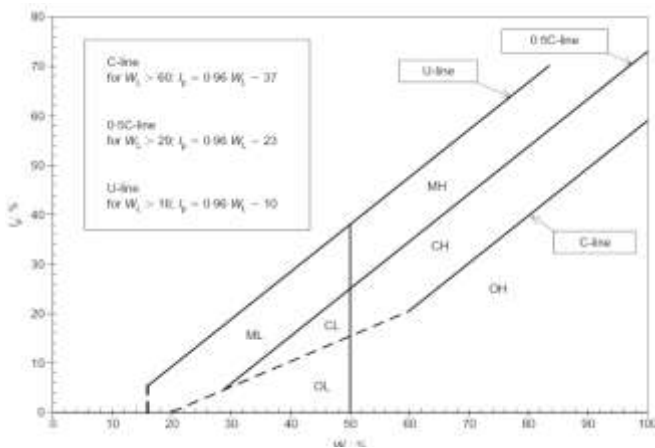


Figure 2 Plasticity chart proposed by E. Polidori (Polidori 2003)

For a given liquid limit, decrease in clay fraction is accompanied by an increase the plasticity index as it requires more plastic clay to maintain the liquid limit at a lower clay proportion. As a result, clays are located below the 0.5C-line and silts above it in the plasticity chart. The dotted line represents the boundary for behavior of inorganic soils.

Organic soils are located below the C-line. In a more recent paper, the author has further expanded on the classification to include coarse grained materials. In this scheme the soils are grouped in to four classes of G Grainy (non-plastic soils), S-G Semi-grainy (mostly non-plastic soils), S-F Semi-fine (plastic soils) and F Fine (plastic soils). Each group is classed based on clay fraction they

contain. Within each group there are symbols used to designate the principal and secondary constituents based on particle size distribution (Ennio Polidori 2015).

Table 1 Summary of classification system proposed by Polidori (Ennio Polidori 2015)

Criterion	Soil Group	Clay Fraction (%)	Principal Soil Component	Second Component
Soil behavior dominated by granular phase characteristics	G Grainy (non-plastic soils)	< 10	Gravel > 2-63 mm	Gr sa, si, (cl)
			Sand > 63µm-2mm	Sa gr, si, (cl)
			Silt > 2µ-63µm	Si sa, cl, gr
Transitional behavior from grainy to fine	S-G Semi-grainy (mostly non-plastic soils)	10-30	Gravel > 2-63 mm	Gr sa, si, (cl)
			Sand > 63µm-2mm	Sa gr, si, (cl)
			Silt > 2µ-63µm	Si sa, cl, gr
Soils as CF increases	S-F Semi-fine (plastic soils)	31-50	Clay < 2µ	Cl (si, (sa), (gr))
			Gravel > 2-63 mm	Gr (cl), (si), (sa)
			Sand > 63µm-2mm	Sa cl, (si)
Soil behavior dominated by the clay-water system	F Fine (plastic soils)	> 50	Silt > 2µm-63µm	Si cl, (sa)
			Clay < 2µm	Cl si, sa, gr
			Clay < 2µm	Cl

José Manuel Moreno-Maroto and Jacinto Alonso-Azcárate (2018) have also proposed and a new plasticity chart and a new definition of clays. The proposed system relies on maximum toughness as a quantitative parameter in the delineation of clays with silts along with the liquid limit and the plasticity index. Measures of maximum toughness are based on a modified rolling apparatus developed by G. E. Barnes (2009). The maximum toughness can be viewed as the maximum resistance, measured in energy per unit volume of soil, to deformation offered by the soils while still remaining plastic (Moreno-Maroto and Alonso-Azcárate 2018; Barnes 2009).

A correlation of the maximum toughness ( $T_{max}$ ) with the plasticity index to liquid limit ratio (PI/LL) was used by Moreno-Maroto and Alonso-Azcárate. The authors redefined clay as having a maximum

toughness of at least  $20 \text{ KJ/m}^3$  and based on the correlation this corresponds to  $PI/LL \geq 0.4937$ . In addition, the lower limit of  $T_{max} = 0$  corresponds to  $PI/LL = 0.3397$ . This data was used to plot the new plasticity chart (see Figure 8).

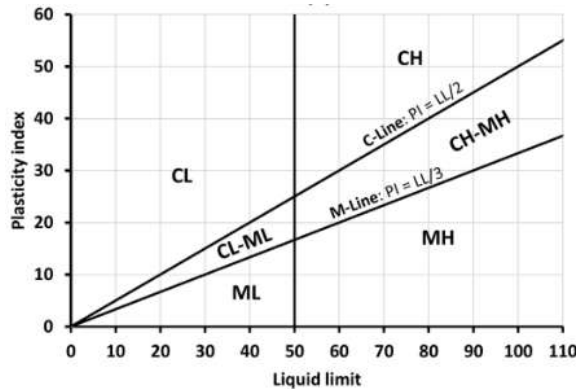


Figure 3 Alternative plasticity chart proposed by Moreno-Maroto and Alonso-Azcárate (Moreno-Maroto and Alonso-Azcárate 2018)

### APPROACH

For this study, representative disturbed soil samples were collected from the Tulu-Dimtu area in the outskirts of Addis Ababa from two adjacent open pits at depths ranging from 1.20m to 3.00m. The Tulu-Dimtu area is selected as there is previous known investigation conducted by Best Consulting Engineers PLC that has resulted in the classification of Addis Ababa black clay as an elastic silt.

The samples were distributed to two laboratories, Best Consulting Engineer's laboratory and Ethiopian institute of Architecture, Building Construction and City development's Material Research and Testing Centre.

As stated previously testing was conducted per ASTM methods but in the hopes of understanding the reason for the stated departure, certain controlling parameters were varied.

The parameters selected are those believed to cause change in the manner in which Addis Ababa black clays are classified, these include:

1. Sample preparation methods,
2. Purity of water used for liquid limit and plastic limit testing,
3. Experience level of individuals performing Atterberg limit tests
4. Variation between laboratories

To assess the effect of sample preparation, two oven dry specimen and two wet-prepared specimens per ASTM D421 and ASTM D2217, respectively, were prepared at the Materials Research and Testing Centre and tested accordingly. Further from each of the pair tested, one in each experiment group (dry prepared or wet prepared) were tested by a laboratory technician while the other by the researchers. This is used to assess the influence of experience level.

To investigate the influence of the purity of water may have on Atterberg limits, one sample was tested using tap water while another using distilled water at Best consulting and Engineers laboratory. The results collected were analyzed to investigate the influence each parameter has in engineering classification of Addis Ababa black clay and conclusions made.

### DATA COLLECTION & ANALYSIS

#### Test Pits

Two test pits were located as show in the map in Figure 10. The pits are excavation pit dug for building construction purposes. The pits had been dug to a depth of more than 3.00m. The soil observed in both pits was black in color with white nodules, had no odor, was wet, had a soft consistency, very high plasticity, no response to dilatancy and when dry they had very high strength,

further, slicken sides were also observed in the cleaved soil blocks.

The while nodules were approximately coarse sand to gravel sized. The nodules could be scratched by fingernails but due to the lack of hydrochloric acid the calcium carbonate presence could not be detected.



Figure 4 Topographic map of Tulu-Dimtu developed using Google Earth and Global Mapper

Three samples were collected from the Tulu-Dimtu area from two pits located relatively close to each other.

Table 2 Location of test pits

Test Pits	Coordinates in UTM		Ground Level Elevation (m)
	Latitude	Longitude	
TP-1	8°53'18.02"N	38°48'50.31"E	2164
TP-2	8°53'20.60"N	38°48'49.72"E	2168

Two samples at depths of 1.20-1.50m and 2.80-3.00m from the natural ground level were collected from TP-1; for the sake of simplicity this samples are designated as S-1 and S-2, respectively. One sample from TP-2 at a depth of 1.50-1.65m from the natural ground level was collected. This sample is designated as S-3. All samples collected were disturbed samples. The samples were manually dug using a pick and a shovel. The collected samples were properly labeled, double packed in common polyethylene bags and transported. Sample S-1 was

transported to the Materials Research and Testing Centre while samples S-2 and S-3 were transported to Best Consulting Engineer's laboratory.

### Laboratory Experimentation

The laboratory tests conducted for the purpose of this study are those relating to engineering classification of soil, this are particle size analysis, liquid limit, plastic limit and specific gravity. Specific gravity is not used in engineering classification of soils, but it is an input in sedimentation analysis of the particle size analysis. In the laboratory experimentation four relevant parameters were varied in the hopes of better understanding the reason for the departure for the comely accepted classification of Addis Ababa black clay.

The four parameters selected are:

1. Sample preparation methods,
2. Purity of water used for liquid limit and plastic limit testing,
3. Experience level of individuals performing Atterberg limit tests and,
4. Variation between laboratories

### RESULTS AND DISCUSSIONS

Classification of the specimen tested is conducted as per ASTM D 2487's unified soil classification system.

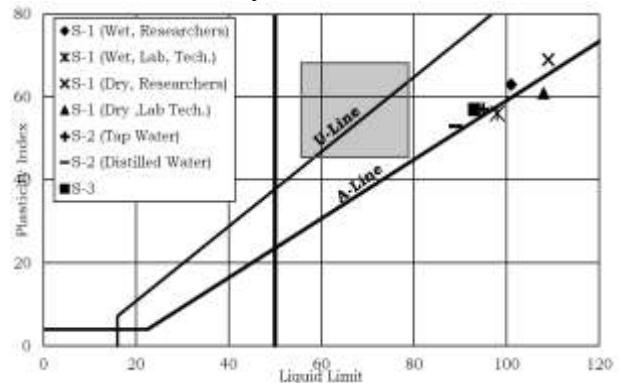


Figure 5 Casagrande plasticity chart with test data plotted

### **Effects of variations in laboratories and operators on index tests and classification**

From the collected data it is noted that there is some variation in all the three index properties measured in the two laboratories and by different operators. To investigate the effect of changes between laboratories and operators may have on index properties and classification, test results compared.

The tests compare are those conducted under similar sample preparation, apparatus and reagent. In addition, variation in results may result from the inherent variation in the soil or from extraneous variables that could not be controlled or were not controlled during the testing. In comparing operators, it is assumed that all the researchers have equal experience levels. This assumption is reasonable as all the researchers have limited testing experience. The specific gravity tests were all conducted by the researchers. Apart from sample S-1, the rest were determined using the dry method. The specific gravity of specimen by the dry method presents a small variation in results the deviation being 0.01. This is in line with ASTM D 854 acceptable range for multi-laboratory reproducibility.

It may be hypothesized here that as the specific gravity has very little operator dependence, large variation is not expected if testing is conducted in line with equivalent standards. Regarding the particle size analysis, all tests were conducted by the researchers. It is observed that variation in results changes with particle size. Relatively small variation is observed in the coarse-grained fraction with the maximum difference of 2.68% from the No. 200 sieve between sample S-1 (dry prepared) and S-2. In the fine-grained fraction variation is seen to increase with reduction in particle size with the largest difference of 8.34, occurring at a particle size of approximately 0.001mm. This comparison is made between the dry

approaches. As the dry approach involves the utilization of pulverization techniques, the force required for breaking the aggregates may vary between laboratories and individuals preparing the samples. Such differences may result in grain size, if excessive force is used to break the aggregation resulting in fracturing the particles. This is especially true for Addis Ababa black clay which has high dry strength and thus requiring considerable pulverization effort. In considering the liquid limit, comparison of sample S-1 (dry, conducted by the researchers using tap water at MRTC) and Sample S-2 (dry, conducted by the researchers using tap water at Best consult), it is observed that there is a difference of 16. This comparison is made assuming the chemical makeup of the water supplied to MRTC and Best consult are identical.

This value is more than the ASTM D4318 value for multi-laboratory reproducibility of high plastic soil by an amount of 12. The ASTM D4318 value for multi-laboratory reproducibility of high plastic soil is obtained ensuring tests are conducted according to the standard. In this regard, the tests deviated from the standard in that they employed tap water. It should also be noted that wear and tare of equipment may also have a part to play. This excessive difference is an indication of the sensitivity of the liquid limit to operator and laboratory variations and the simple following of standardized test procedures alone is not adequate to obtain precise results.

The liquid limit conducted at the Material Research and Testing Centre were also conducted by a laboratory technician. Comparing Samples S-1(wet, conducted by a laboratory technician) with S-1(wet, conducted by researchers) and S-1(dry, conducted by a laboratory technician) with S-1(dry, conducted by researchers).

In addition, assuming all the researchers have equal experience. It is observed that the differences are 1 and 3 for the dry and wet prepared methods, respectively. This variation is within ASTM D 4318 range of acceptable multi-laboratory test results. In contrast to the variation between laboratories these values are small. As the researchers are not complete novices but at least understand, theoretically, the determination of the liquid limit, it may be inferred that variations in operator are minimal, provided that the operator has some limited knowledge of the test. Regarding the plastic limit the multi-laboratory variation of the mean of two PL test runs between dry prepared samples is 4. This variation is within ASTM D 4318 range of acceptable multi-laboratory test results.

It should be noted here the variation within each test and between test runs is larger, the greatest being 10. As two test runs of a given test are conducted by the same researcher in successive order, it shows the lack of repeatability in the test. When evaluating the differences in result between the researchers and the laboratory technician, it is observed that a difference of 4 and 7 for the wet and dry method, respectively.

The differences in between test runs show that the largest difference for the test conducted by the technician is 6 while it is 10 for the researchers. This again shows the lack of repeatability in the test while the repeatability of the lab technician is better it is still high. As is expected the repeatability is dependent up on experience level. In the variation of the plasticity index, which is the result of the propagation of the variations in the LL and PL? It is observed that maximum multi-laboratory variation for the dry method is 12 which is considerably larger than the ASTM D 4318 acceptable value. The deviation between researchers is 8 and 7 for the dry method and wet method, respectively. In the classification of the soil,

which ultimately relied on the plasticity chart, which in turn relies on the LL and PI, it observed that while all samples are located close to the A-line. Two of the samples, S-1, conducted by the laboratory technician using dry and wet method plotted below the A-line. Thus, in cases where the sample plots close to the A-line, it is possible for variations in operator and laboratory to cause misclassification.

### **Effect of sample preparation on liquid limit, plastic limit, and classification**

Two sample preparation techniques were used dry method and the wet method according to ASTM D422 and D2218, respectively. Comparison is made between tests results from samples prepared by the two methods in the same laboratory and tested by similar operators. Regarding the liquid limit the dry method of sample preparation resulted in a higher value than the wet method when both researcher and lab technicians conducted the test. This is also true for the plastic limit as well. This clearly indicates that oven drying, and pulverization have an influence on the liquid and plastic limit. In regard to soil classification test conducted by the researchers resulted in CH classification in both wet and dry prepared samples while MH classification was obtained when laboratory technician conducted the tests.

### **Effect of water chemistry on liquid limit, plastic limit and classification**

To investigate the effect of water chemistry on liquid limit and plastic limit comparison is made between Atterberg limit tests conducted with distilled water and Tap water in the same laboratory, using the same sample preparation schemes and with the same researcher.



For both the liquid and plastic limit, the utilization of distilled water reduced the test results. This is an indication that the utilization of Tap water can have an effect on the Atterberg limits. In The classification of soil both specimen plotted above the A-line. It should be noted that simply because there was no change in the classification does not imply that water chemistry does not affect soil classification but in this case the tap water used was of adequate quality not to cause changes in classification this may not be always true.

### Make ‘False Elastic’ ‘Fat’ Again

Based on the literature review conducted and based on the limited data from tests, it is possible for Addis Ababa black clay to plot below the A-line in the MH region. But it should be noted, even though the soil may plot below the A-line, it still remains close to it. Based on the conducted experiment it is also possible that operator error may shift the result to CH or MH from the true class. It is further important to point out that such distinction of boundary soil is technical and in practice it is imperative that engineering judgment be employed in interpreting such boundary soil classes. To overcome such irregularities, there are alternative plasticity charts proposed, discussed in the literature review.

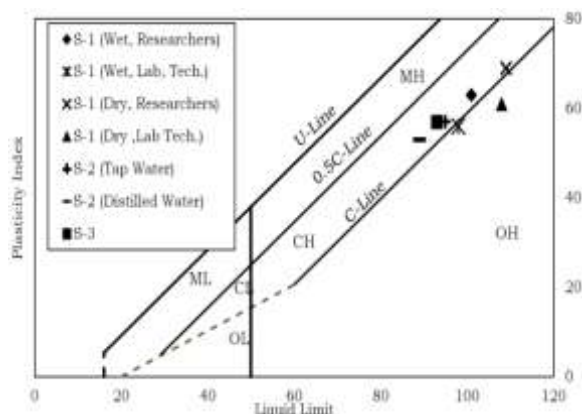


Figure 6 Polidori plasticity chart with test data plotted

The Polidori (2003) plasticity chart which is based on clay size fraction classifies some of the specimen as organic soil which is contradictory to the laboratory tests.

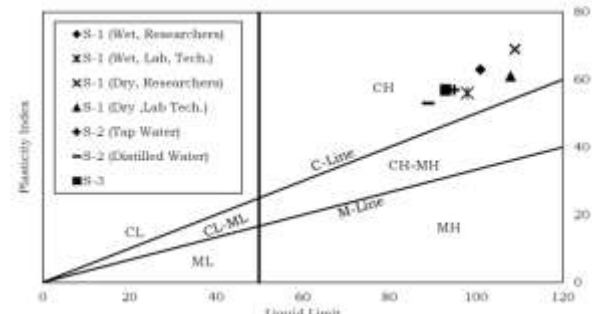


Figure 7 Moreno-Maroto and Alonso-Azcárate plasticity chart with test data plotted

The (Moreno-Maroto and Alonso-Azcárate (2018) plasticity chart which is based on the toughness definition of clays classifies the soil as fat clays. In contrast this chart is more representative of observed phenomenon.

### CONCLUSIONS and RECOMMENDATIONS

Based on survey of literature and the limited laboratory tests conducted, it can be stated with a reasonable degree of certainty that Addis Ababa black clays may plot below the A-line. But it should be noted that they remain close to the A-line. From laboratory test conducted it can be stated that test used to determine index properties for classification are dependent on experience of operator and variations between laboratories. This is true especially for Atterberg limits. Due to this dependency and due to the fact that the plot is located at the boundary region of the A-line, it is possible for the plot of the plasticity data to plot on either side of the A-line while its true location is on the other side.

This results in a misclassification. Classification tests are also dependent on sample preparation. Samples prepared by the dry method have higher liquid and plastic limit than those determined by the wet



method. In the tests conducted this variation did not cause change in the classification of the soil. A fourth factor considered was the utilization of tap water versus distilled water in the determination of Atterberg limits. It was found that tap water resulted in a higher liquid and plastic limit than distilled water. In the tests conducted this variation did not cause change in the classification of the soil. It should be noted that apart from the possibility of Addis Ababa black clays being plotted below the A-line which is backed by literature survey, the remaining assertion made require further investigation as only limited laboratory testing was done.

Based on the above discussion the following recommendations are given regarding laboratory testing for the determination of index properties and engineering soil classification:

1. Simply following of standardized testing procedures is not enough to ensure accuracy and precision of laboratory testing, it is necessary to regularly maintain and calibrate laboratory equipment.
2. As classification tests, especially Atterberg limits, are operator dependent, regulatory bodies should ensure the qualification of laboratory technicians.
3. Regarding Addis Ababa black clays one should utilize the wet method of sample preparation at the very least Method A (air drying) but preferably Method B (washing method) of ASTM D4318. Furthermore, one should use distilled in the determination of the Atterberg limits.

Soil classification is ultimately way of communication and a means of estimating engineering properties. As Addis Ababa black clays exist at the boundary region of the A-line on either side, it important to recognize this characteristic in reporting and interpreting geotechnical investigations. This

study investigated in a limited and general manner the effects of the parameters previously mentioned on engineering soil classification as it applies to Addis Ababa black clays. In future research the effect of each parameters can be investigated in detail and additional factor included.

## REFERENCES

- [1] Al Haj, K.m.a., and J.r. Standing. 2015. "Mechanical Properties of Two Expansive Clay Soils from Sudan." *Géotechnique* 65 (4): 258–73.
- [2] Barnes, G. E. 2009. "An Apparatus for the Plastic Limit and Workability of Soils." *Proceedings of the Institution of Civil Engineers Geotechnical Engineering*.
- [3] Best Consulting Engineers PLC. 2012. "Geotechnical Investigation Report to Addis Ababa Housing Development Agency." Addis Ababa, Ethiopia: Best Consulting Engineers PLC.
- [4] Haigh, S.k., P.j. Vardanega, and M.d. Bolton. 2013. "The Plastic Limit of Clays." *Géotechnique* 63 (6): 435–40.
- [5] Moreno-Maroto, José Manuel, and Jacinto Alonso-Azcárate. 2018. "What Is Clay? A New Definition of 'Clay' Based on Plasticity and Its Impact on the Most Widespread Soil Classification Systems." *Applied Clay Science* 161 (September): 57–63.
- [6] Morin, W. J., and W. T. Parry. 1971. "Geotechnical Properties of Ethiopian Volcanic Soils." *Geotechnique* 21 (3): 223–32.
- [7] O'Kelly, B. C., P. J. Vardanega, and S. K. Haigh. 2017. "Use of Fall Cones to Determine Atterberg Limits: A Review." *Géotechnique* 68 (10): 843–56.

- [8] Polidori, E. 2003. "Proposal for a New Plasticity Chart." *Géotechnique* 53 (4): 397–406.
- [9] Polidori, Ennio. 2015. "Proposal for a New Classification of Common Inorganic Soils for Engineering Purposes." *Geotechnical and Geological Engineering* 33 (6): 1569–79.
- [10] Sivakumar, V., D. Glynn, P. Cairns, and J.a. Black. 2009. "A New Method of Measuring Plastic Limit of Fine Materials." *Géotechnique* 59 (10): 813–23.
- [11] Smith, C. W., A. Hadas, J. Dan, and H. Koyumdjisky. 1985. "Shrinkage and Atterberg Limits in Relation to Other Properties of Principal Soil Types in Israel." *Geoderma* 35 (1): 47–65.
- [12] Teferra, A., and S. Yohannes. 1986. "Investigations on the Expansive Soils of Addis Ababa." *Zede Journal* 7 (0): 1-9–9.
- [13] Teklu, Daniel. 2003. "Examining the Swelling Pressure of Addis Ababa Expansive Soil." Thesis, Addis Ababa University.



# BACTERIAL CONTAMINATION OF SCHOOL'S DRINKING WATER IN

## ADDIS ABABA, ETHIOPIA

Dawit Debebe<sup>1</sup>, Zerihun Getaneh<sup>1</sup>, and Fiseha Behulu<sup>1\*</sup>

<sup>1</sup> School of Civil and Environmental Engineering; Addis Ababa Institute of Technology (AAiT),  
Addis Ababa University, Ethiopia

\*corresponding author: [fiseha.behulu@aait.edu.et](mailto:fiseha.behulu@aait.edu.et)

### ABSTRACT

Access to safe drinking water and hygienic living conditions is a global concern and these issues are especially serious in developing countries. The objective of this study is to evaluate the quality of water consumed by kindergarten schools' children in Addis Ababa city, who are highly susceptible to issues associated with microbial contamination in water. Total coliforms, *E. coli*, pH and residual chlorine in the water distribution system were measured at three water sources and 38 schools. The microbial analysis result shows 7 out of 38 schools were contaminated with total coliform bacteria. However, *E. coli* was not detected in any of the samples, meaning that all samples were free from fecal contamination. In addition, the free chlorine level of the samples was also tested. The results indicated that 16 out of 38 (42.1%) of the water samples had a free chlorine value below the WHO recommended 0.2mg/L. It is therefore, possible to conclude that the efficiency of a water supply infrastructure determines the concentration levels of microbial contamination and residual chlorine that reaches the end users. The study addresses critical issues and methods to mitigate the problems caused by microbial contamination in water supply distribution infrastructure.

**Key words:** Addis Ababa, *E.coli*, Residual Chlorine, Total Coliforms

### INTRODUCTION

A healthy and safe school environment encompasses the physical surroundings, the psychosocial, learning, and health-promoting environment of the school. Additionally, hygienic practices, such as accessing to sanitation and providing clean water are all important contributors to children's health [1].

Access to clean water and sanitation is declared as a human right by United Nations in 2010. It is a prerequisite for the realization of many human rights, including those relating to people's survival, education and better standard of living. Safe drinking water and hygienic living conditions is a global concern and these issues are especially serious in developing countries, like Ethiopia that have suffered from a lack of safe drinking water and inadequate sanitation services [2]. In several educational institutions, waterborne diseases have become common problems causing health complications on children and adults. This may be related to contamination of water tanks or infiltration of the microorganisms in water pipes. According to a data from the educational statistics annual abstract (2017) taken from Addis Ababa Education Bureau, there are 164,072 students, 51.16% male and 48.84% female, attending in 1172 Kindergarten schools in the city.

All of the children are directly affected by the contamination of water they get from their school's tap water. Their families are also indirectly affected by costs of medical treatment they spend on their children, which is unaffordable to most of the poor family members living in the city. Therefore, the monitoring and further analysis of the quality of water originating from faucets for school children's consumption becomes important for diagnosing the problem and to develop prevention and mitigation strategies.

Microbial contamination is by far the most important public health challenge of drinking water supply systems. All microbial organisms of viral, bacterial, parasitic and protozoan origins can be found in the distribution network of the water supply [3]. These harmful organisms can originate from a variety of sources such as industrial waste, decayed plant matter, agricultural runoff and human wastes. Some of these microbial organisms are more pathogenic than others. The hazardous pathogens in drinking water are usually associated with human or animal excreta in many circumstances, but there are also other pathogens capable of causing infection through the drinking water. The most transmissible diseases related to drinking water are those caused by pathogenic viruses, bacteria and parasites[4]. Examples of pathogenic organisms implicated for water borne disease outbreaks include *E. coli* O157:H7, *Salmonella*, *Norovirus*, *Cryptosporidium* and *Giardia*. These pathogens are also different in characteristics, behavior and resistance. Simultaneously they affect different persons in various ways, reliant on factors as age, sex, state of health and living conditions [4]. This study is focused on indicator organisms (total coliforms and *E. coli*) in characterizing the microbial quality of the water from the distribution network.

Addis Ababa has grown very rapidly since it was founded in 1886. This growth has put enormous pressure on water supply services and the sewerage system. The water supply infrastructure in the city is more than 40 years old and is known for its low output capacity and high-water losses due to degraded pipelines [3]. The water supply infrastructure in the city is more than 40 years old and is known for its low output capacity and high-water losses due to degraded pipelines[3]. Similarly, Abay[5] has also stated that the growth of Addis Ababa City has been unregulated and unstructured and the city has not had formal urban planning until recently. This has put many constraints on the water supply system. A major concern is the significant losses of water caused by leakage from the old supply infrastructure.

The national drinking water standards are identical to the World Health Organization's[4] guideline for the provision of safe drinking water. However, the treated water is generally delivered to households and schools in old metallic (galvanized iron and cast iron) pipelines. Some piping has been replaced by HDPE and PVC materials. Pipes are either buried underground or exposed to the environment. In many of the slum dwellings, the pipelines are very old and degraded. Approximately 30-40% of the drinking water supplied to the city does not reach consumers. The water is lost at different levels of the distribution system due to leaking pipes and aging infrastructures [3, 5].

The combination of the degraded infrastructure and a cross-connected distribution system may provide a favorable environment for drinking water contamination to occur. Considering the poor environmental conditions in many districts of the city, there are many chances for drinking water contamination in cracked and leaky water supply pipes. Currently,

## Bacterial Contamination of School's Drinking Water in Addis Ababa, Ethiopia

there is no comprehensive water quality monitoring or data for drinking water quality at the household and school levels. It is therefore unclear how much contamination is occurring to the drinking water quality once it is distributed from the treatment plants, and whether the water is safe to drink once it reaches school.

### MATERIALS AND METHODS

#### Study area and sampling locations

The study was conducted in Addis Ababa, capital city of Ethiopia, which has a population of more than 4 million in an area of 540 km<sup>2</sup>[6]. The city gets its treated water from three sub-systems:

- A. Akakisubsystem is located in the southern part of the city. It has a groundwater source and its treatment system is mainly limited to disinfection (chlorination).
- B. Legedadi subsystem has both surface and groundwater sources which are situated in the western part of the city.
  - Its surface water source part of this subsystem has a conventional water treatment system. This system includes pre-chlorination, coagulation, sedimentation, filtration and post chlorination components [7].
  - The groundwater source from Legedadi subsystem has a treatment system of disinfection (chlorination). These two systems then blended at some central reservoirs for further distribution to the end users.
- C. Gefersa subsystem is located in the northwestern end of the city. It has surface water source with conventional water treatment system. This system is the same as Legedadi (surface water source) and it includes pre chlorination,

coagulation, sedimentation, filtration and post chlorination [7].

According to Addis Ababa Education Bureau there are 164,072 kindergarten children in the city. For every 4000 children one representative water sample was collected. Therefore, a total of 41 samples are needed. But since the city has a problem regarding shortage of water, thirty-eight samples were collected. Fifteen kindergarten schools from *Akakissub*-system, fifteen from *Legedadi*, and eight schools from *Gefersa* sub-systems were selected according to the sub system's coverage areas. Random sampling technique was used to select the schools in the sub-systems. But the schools were chosen in a way that they would be representative of their sub-system as shown in figure 1. The distance from the schools to the treatment plants can also be seen in figure 1. The samples were collected from 11 May 2018 to 17 May 2018 on a dry season. The sampling was carried out based on the standardized sampling techniques as outlined in USEPA guidelines for water testing [8].

One water sample was taken from each school giving a total of 38 samples. However, from the sources, two water samples were taken from each treatment plant, before and after the water is treated, which means six samples have been taken from the three treatment plants. The total number of samples taken is sum up to 44. The samples from the schools were taken from a tap which was directly connected to the municipal water supply. Flushed water samples were taken and each sample had a volume of 500-1000ml, collected using pre-labeled 500 -1000 ml sterile plastic bottles. The bottles were initially cleaned using standard detergents and distilled water. The water samples were transported to the Addis Ababa University Faculty of Science, Department of Microbiology laboratory. The samples were then tested for pH (measured



onsite), residual chlorine, E-coli and total coliforms within 24 hours of sampling. Results on the blood lead level from the same sample points were given in previous work by Debebeet. al. [9].

### **Measuring chlorine**

The residual chlorine in the water was tested using portable Palintest 7100 photometer (i.e. made in England for water quality and wastewater tests). The instrument has dual light source photometer offering direct reading of pre-programmed test calibrations, absorbance and transmittance. It works in wave length ranges of 450nm, 500nm, 550nm, 570nm, 600nm, and 650nm at measurement accuracy of  $\pm 1\%$ . During the test, a reagent called diethyl-p-phenylene diamine (DPD1) was used. DPD1 reacts with chlorine in water and changes its color to pink. The change in color is read by the photometer to get the residual chlorine content of the sample water.

### **Measuring pH and Temperature**

pH and temperature were measured simultaneously using a hand pH meter. Each sample was poured in a beaker and the hand pH meter was inserted. Each sample was measured 3 times and an average result was taken.

### **Microbial Analysis for Total Coliform and E. coli**

Total coliform counts were carried out by membrane filtration technique[10]. A sterilized pad dispenser was used to introduce the growth absorbent pads into the base of Petri dishes, and the growth pads were saturated with the Lauryl Sulphate Broth. 100ml water sample was filtered using a membrane filter ( $0.45\mu\text{m}$ ) in a vacuum filtration apparatus, and all the filters were transferred to the absorbent pad which was saturated with the broth.

The Petri dishes were incubated at  $37^{\circ}\text{C}$  for 4hr for resuscitation to recover physiologically stressed coliforms before incubation. Then after, plates for total coliform counts were incubated at  $37^{\circ}\text{C}$  for 24hrs, and then colonies were counted and recorded.

*E. coli* was tested using Eosine Methylene Blue (EMB) agar. This selective media grows only gram-negative bacteria. Since *E. coli* are groups of gram-negative bacteria, it was possible to test using this media. If *E. coli* bacteria are present in the sample, it shows a metallic green color on the media after it's kept in an incubator for 24 hours at  $37^{\circ}\text{C}$  [11].

The samples were carefully processed in FASTER TWO 30hub. This hub creates a vertical laminar flow which guarantees excellent decontaminated working area and particle-free conditions. Also, to prevent any environmental contamination, the media and petri dishes were autoclaved. The researcher's hands were also sanitized with 70% denatured ethanol at all times during the work on the hub to prevent contamination. The processed samples were finally put in an oven for 24 hours at  $37^{\circ}\text{C}$ .

As a quality control mechanism, all sampling bottles were appropriately labeled, and the samples were collected using standardized drinking water sampling techniques. The collected water samples kept in icebox during transportation put at 4 degree Celsius before analysis in the laboratory.

Before analysis, sterilization of required laboratorial equipment and culture medium was carried out. Moreover, to ensure the validity of the analysis, blank samples were analyzed following the same procedure. Water quality analysis guideline, protocol, and quality control were used.

# Bacterial Contamination of School's Drinking Water in Addis Ababa, Ethiopia

## RESULTS AND DISCUSSIONS

In Addis Ababa rapid urbanization and population growth are taking place. This rapid growth has led to an increasing demand for water which is growing at a faster rate than the supply. Even though Addis Ababa Water Supply and Sewerage Authority (AAWSA) is working to increase the supply capacity, it is currently not able to supply enough drinking water to the growing population. This has resulted in water shortages in many areas of the city. As a result, drinking water is now being supplied in an intermittent manner. Unscheduled water supply disruptions are common in many parts of the city. It is common for tap water to be supplied only once per week in some parts of the city. This is worst for residents located at higher altitudes and those living in the higher floors of condominium apartments. Such challenges are further resulted in insufficient pressure in the system to supply the water to elevated areas unless a booster pump is used. The combination of scheduled water supply and an old, leaky distribution systems result in low pressures in the distribution network. This can result in the intrusion of external contaminants into the leaky and cross-connected infrastructure during supply interruption and reinstatement events.

### pH and Temperature

The results showed that the average temperature records of water samples taken from the schools was 25.4°C ranging from 22.4°C to 28.1°C. Similarly, earlier studies in Gondar zone [12], Bahir Dar [13] and Nekemt[14] reported a mean temperature of 21.3°C, 23.8°C and 20.8°C, respectively. In tropics, the climate is characterized by high

temperature and convective rainfall, and these factors might have contributed to the high temperature records of water samples from different cities of Ethiopia that did not meet the WHO standard of < 15°C [14].

*Akaki* catchment having a ground water source has the largest mean and median PH values of 7.96 and 8 respectively. The next is the *Gefersa* catchment with mean and median PH values of 7.75 and 7.71 respectively. Finally, *Legedadi* catchment has the lowest mean and median PH values of 7.6 and 7.57. PH results for the samples taken from the treatment plants is given in table 1 and PH values of the schools is given in figure 1.

Table 1 PH levels of samples taken from the treatment plants

Sources	pH	
	Before Treatment	After Treatment
Akaki treatment	8.33	8.2
Legedadi treatment	7.9	7.7
Gefersa treatment	7.3	7.72

For comparison, the average PH levels of various cities' water sources are given in the table below. The variation could be due to geological conditions of the water sources.

Table 2 The average PH levels of various cities' water sources

City	PH level	Reference
Ziway	8.3	[15]
Adama	7.8	[16]
Nekemte	6.8	[14]

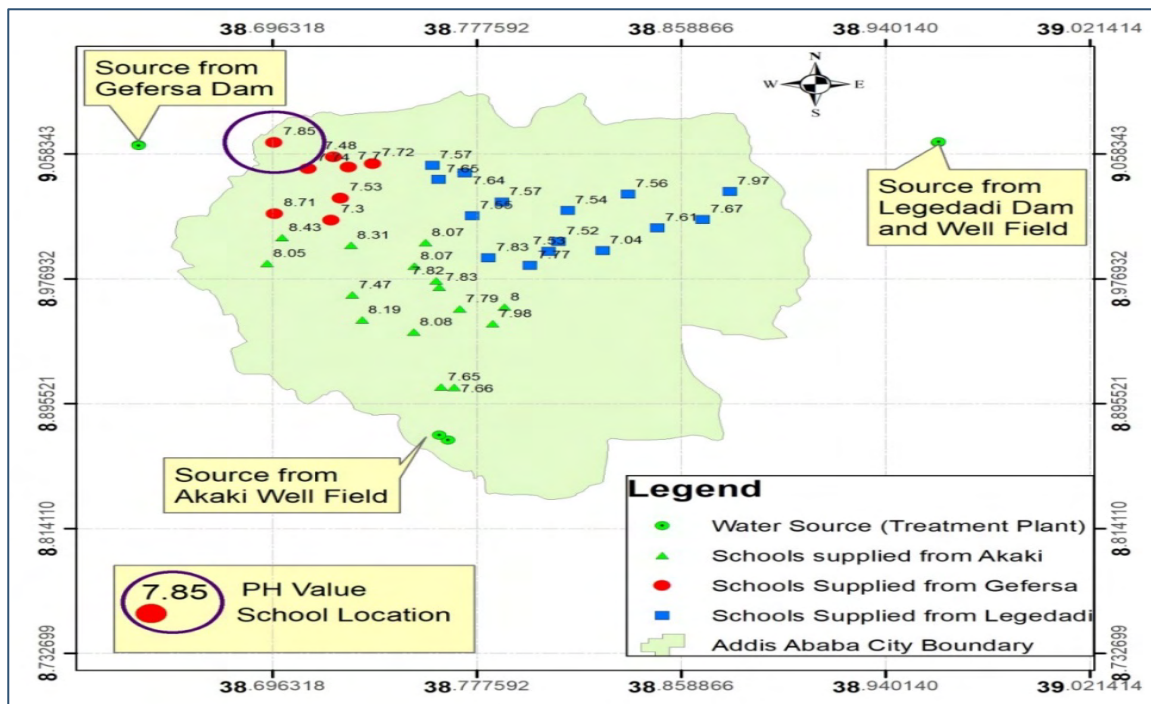


Figure 1 Distribution of PH in Addis Ababa

All samples remained within the recommended standard limits of 6.5-8.5 as noted by WHO [17] and ESA [18]. The average pH levels from all 38 schools were 7.77 and the pH levels measured from the schools' tap water were generally lower than the source for all sub-systems. Both median and mean values of the samples from the schools were also smaller than the source water (AAT, LAT and GAT). The slight reduction in the pH values measured in the water samples may be attributed to the corrosion of aged and cross-connected metallic pipeline materials used in the water supply distribution system. This decrease in pH is consistent with the study by Mekonnen [3] who reported that the pH in drinking water decreases as a result of

corrosion taking place in distribution systems.

### Free Chlorine

The minimum recommended WHO value for free chlorine residue in treated drinking water is 0.2 mg/L. In this study, 16 out of 38 (42.1%) of the school water samples had a free chlorine value below 0.2 mg/L. Highlighted samples in figure 2 show samples having free chlorine level below 0.2 mg/l. Similar studies showed that 15.2%, 37.5%, and 95.7% of tap water samples from tap water distribution systems in Nekemte [14], Ziway [15] and Bahr Dar towns [13] contained lower free chlorine than the recommended limits.

## Bacterial Contamination of School's Drinking Water in Addis Ababa, Ethiopia

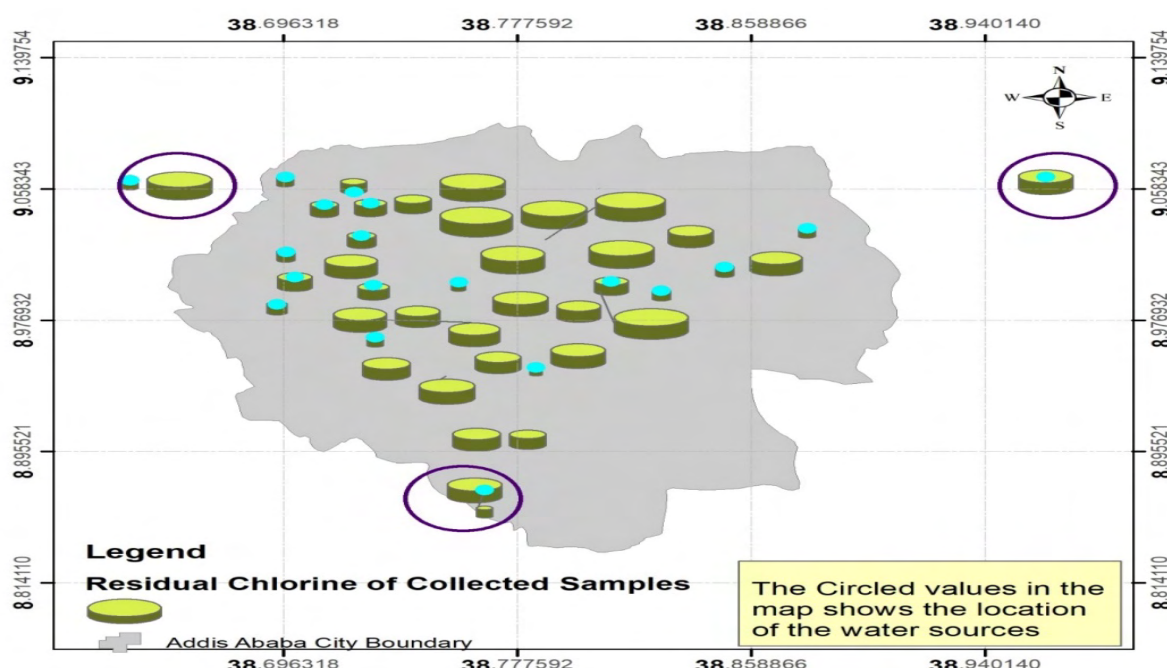


Figure 2 Free chlorine distribution (highlighted samples have residual chlorine below 0.2mg/L)

In the *Akaki* catchment, 40% (6 out of 15) of the samples have residual chlorine values below the recommended 0.2mg/L. For *Legedadi*, 26.67% (4 out of 15) and for *Gefersa* 75% (6 out of 8) of the samples have values below 0.2mg/L. The free chlorine levels of the samples from the treatment plants are also listed in table 3.

Table 3 Free chlorine levels of samples taken from the treatment plants

Sources	Free Chlorine (mg/l)	
	Before Treatment	After Treatment
Akaki treatment	0.04	0.45
Legedadi treatment	0	0.45
Gefersa treatment	0	0.04

Similar studies showed that the free chlorine level of water samples from disinfection point in Nekemte town was 0.23mg/l[14]. The treatment outlet of Ziway town had free chlorine of 0.79mg/l[15]. But unlike these two

studies, 0.03mg/l free chlorine was recorded from the main distribution tank of Bahir Dar town[13] which is similar to the free chlorine level of Gefersa treatment plant as seen on table 2.

For the treatment plant assessment, chlorine residue was tested based on the data collected on the 18th of May 2018. The test results revealed that treated water leaving Gefersa treatment plant had no residual chlorine. Since this was not logical, that water leaving a treatment plant must have residual chlorine, another sample was collected on the 19<sup>th</sup> of May 2018 in order to clarify such issues. But the sample from had a value of 0.04mg/l which was less than the WHO recommended 0.2mg/l. This clearly shows the poor management and quality control works in the treatment plants.

For the assessment of distribution systems' performance in terms of residual chlorine, it is expected that the concentration degrades when treated water enters into the distribution system. A possible reason for

this rapid drop in concentration could be due to the breakdown of residual chlorine by microbes attached to biofilms, corrosion in pipes and water aging in distribution system. Another possible reason could be the intermittent supply of water that can lead to negative pipe-pressure and intrusion of contaminants. These contaminants could further reduce the residual chlorine in the distribution system. The distance of the schools to the treatment plants and increasing time spent in water storage reservoirs and pipes could also deplete the residual chlorine before it reaches the schools taps. These assumptions are similar to study findings by Mekonnen[3] who reported that rapid deterioration of residual chlorine occurred in the water distribution network of *Legedadi* sub-system. This was a result of, the distance from the treatment plant, the intermittent supply leading to contaminant intrusion, and growth of bacteria in pipes due to the depletion free residual chlorine. In addition, a study by Kumpel and Nelson [19] compared the microbial water quality in an intermittent and continuous piped water supply. It was reported that a significantly higher proportion of samples collected from a continuous supply met the minimum standard for residual chlorine concentrations when compared to samples from intermittent water supplies.

### Microbial Analysis

Bacteriological analysis of the samples revealed that there was total coliform bacteria contamination in the three catchments. Accordingly, 3 out of 15 samples from *Akaki* ,2 out of 15 from *Legedadi* and 2 out of 8 samples from *Gefersa* catchment were contaminated. Table 3 shows the bacterial count for the 3 catchments. Treated water samples taken from the treatment plants were also free from contamination.

Table 4 Total coliform count of samples contaminated

Catchment	Sample Local Name	CFU/100 ml
Akaki	Auxilium catholic school	2
Akaki	Great Ethiopian transformers school	120
Akaki	TibebGebeya school	1
Legedadi	Goro primary school	65
Legedadi	Vision academy	20
Gefersa	BiruhTesfa kindergarten	1
Gefersa	Amigonian school	1

The highest total coliform count was recorded from tap water at Great Ethiopian transformers school in Akaki catchment with 120 CFU/100ml. This is similar with a study by Duressa et. al [14] in Nekemte town that reported a highest total coliform count of 95 CFU/100ml.

All of the samples did not show any sign of *E. coli* contamination. This means the water is safe from fecal contamination. However, on the contrary to the presented findings, a study by Mekonnen[3] showed *E.coli* contamination in *Legedadi* sub-system. This is presumably due to the fact that samples in that study were taken from July to September on the rainy seasons. Therefore, contaminants can easily enter the distribution system in these wet seasons. On a similar study in Nekemte town, 37% of samples showed fecal contamination [14].

Total coliform contamination was found in all catchments and all contaminations are directly related with the free chlorine. The seven contaminated samples had residual chlorine below 0.12mg/l which contradicts

## Bacterial Contamination of School's Drinking Water in Addis Ababa, Ethiopia

the WHO recommended value of 0.2mg/l. Similarly, studies by Kumpel and Nelson [19] in Hubli-Dharwad, India and Mekonnen[3] in Addis Ababa, Ethiopia have also reported frequent and elevated bacterial contamination in tap water samples with residual chlorine concentrations below the recommended guideline values. Zero bacteria counts are reported in water samples retaining good residual chlorine concentrations in both studies which resembles the present study.

The microbial water quality results measured in this study strongly agree with a study conducted by Kumpel & Nelson [19]. It was reported that bacterial contamination is more frequent in intermittent water supply networks when compared to those continuously supplied. The study reported by Mekonnen[3] also suggests that bacterial contamination in an intermittent water supply could be caused to the intrusion of contaminants from the environment when the water supply to pipelines is turned off. This causes negative pipe-pressure events and causes problems when combined with cross-connection pipelines. These issues are common in Addis Ababa and are the main problems within the study areas.

### CONCLUSIONS

Based on the results from this study, we can conclude that the main cause of water quality degradation in the distribution system is likely due to the efficiency of water distribution infrastructures. This is associated with the disruption in water supply, intermittent supply, lack of continuous flow in distribution network and age of pipes which are susceptible to leakages. This results in the intrusion of external contaminants in the pipelines of the distribution system. This may ultimately result in non-compliance with the WHO drinking water standards. To combat such

challenges improving water distribution system efficiency, regular monitoring of water quality level at the source as well as within distribution network and automated system management strategies are relevant recommendations. From school children health point of view localized water supply treatment at school inlet systems are also possible options.

### REFERENCES

- [1] Bakır B, Babayigit MA, Tekbaş ÖF, Oğur R, Kılıç A, Ulus S. Assessment of Drinking Water Quality in Public Primary Schools in a Metropolitan Area in Ankara, Turkey International Journal of Health Sciences and Research. 2015;5(4):257-66.
- [2] Amenu D, Menkir, S., Gobena, T., . Bacteriological quality of drinking water sources in rural communities of Dire Dawa Administrative council. Int J Res Dev Pharm L Sci, . 2013:775-80.
- [3] Mekonnen D.K. The Effect of Distribution Systems on Household Drinking Water Quality in Addis Ababa, Ethiopia, and Christchurch, New Zealand. [MSc. Thesis, ] 2015.
- [4] WHO. Guidelines for drinking-water quality, fourth edition incorporating the first addendum: World Health Organization;; 2017.
- [5] Abay GK. The impact of low-cost sanitation on groundwater contamination in the city of Addis Ababa. [ MSc. Thesis ] 2010.
- [6] CSA. Population and housing census of Ethiopia 2007.

- [7] AAWSA. Annual report of planning, implementation and evaluation of Addis Ababa Water and Sewerage Authority. *Official Report* 2014.
- [8] USEPA. Human Health Evaluation Manual, Supplemental Guidance: Standard Default Exposure Factors 1991.
- [9] Debebe D, Behulu F, Getaneh Z. Predicting children's blood lead levels from exposure to school drinking water in Addis Ababa, Ethiopia. *Journal of Water and Health*. 2020.
- [10] Bartram, Jamie, Ballance, Richard. World Health Organization & United Nations Environment Programme. Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programs / edited by Jamie Bartram and Richard Ballance. London : E & FN Spon. <https://apps.who.int/iris/handle/10665/41851> 1996.
- [11] MacFaddin J.F. Biochemical Tests for the Identification of Medical Bacteria. Baltimore (Md.) PA, USA: Williams & Wilkins Publishers, ISBN-0683053183, 3rd edition; 2000.
- [12] Damite D, Endris M., Tefera Y. Assessment of microbial and physico-chemical quality of drinking water in North Gondar Zone, Northwest Ethiopia. *Journal of Environmental and Occupational Science*. 2014;3(4):170.
- [13] Kassahun G. Physicochemical and bacteriological drinking water quality assessment of Bahirdar town water supply from source to yard connection, North western Ethiopia. [MSc. Thesis]: Addis Ababa University, Ethiopia; 2008.
- [14] Duressa G., Assefa F., Jida M. Assessment of Bacteriological and Physicochemical Quality of Drinking Water from Source to Household Tap Connection in Nekemte, Oromia, Ethiopia. *Journal of Environmental and Public Health*. 2019.
- [15] Bedane K. Assessment of physicochemical and bacteriological quality of drinking water in Central Rift Valley System, Ziway town, Oromia regional state. [MSc. Thesis]. Addis Ababa: Addis Ababa University, Ethiopia; 2008.
- [16] Eliku T., Sulaiman H. Assessment of physico-chemical and bacteriological quality of drinking water at sources and household in Adama Town, Oromia, Ethiopia. *African Journal of Environmental Science and Technology*. 2009;9(5):413–9.
- [17] WHO. Guidelines for Drinking-Water Quality, Surveillance and Control of Community Supplies 1997; volume 3, 2nd edition.
- [18] ESA. Drinking Water Specifications, Compulsory Ethiopian Standards, CE58, (Ethiopian Standards Agency), Addis Ababa, Ethiopia, 1st edition. Official report. 2013.
- [19] Kumpel E, Nelson KL. Comparing microbial water quality in an intermittent and continuous piped water supply. . *Water research*. 2013;47(14):5176-88.



# ANALYSIS OF TRACK GEOMETRY INDEX MEASUREMENT METHODS

Daniel H/Michael<sup>1</sup>, Elias Kassa<sup>1</sup>, Getu Segni<sup>1</sup>

School of Civil and Environmental Engineering, African Railway Center of Excellence

Corresponding author's email: [dh5045@gmail.com](mailto:dh5045@gmail.com)

## ABSTRACT

*Degradation of railway track can be described by main geometry parameters such as profile, alignment, gauge, cant, and twist but track geometry quality index can be used for aggregating two or more geometric defects and represent health condition of track structure. This paper discusses different methods of quality indexes and analyzes numerically three methods based on real track geometry measurement data from Addis Ababa – Djibouti railway line and their advantages discussed for the purpose of recommending TQI method for predicting future state of track which will be used in Predictive maintenance. Data collected is from 25-27 of May 2020 for 215.8Km length. Results from analysis shows, track geometry index (TGI) represents track quality more reasonably. Chinese TQI method can also represent track quality but gives equal weightage for all types of degradation parameters on the other hand TGI allocated more weightage for parameters with higher effect on ride quality. J synthetic method can only represent two types of quality below and above threshold but the two other methods represent more quality levels. Theoretically, advantages and disadvantages of methods discussed can be referred but practically recommended method can be used in prediction models for implementing predictive maintenance.*

**Keywords:** Track degradation, Track geometry defects, track quality Index, track quality level,

## 1. INTRODUCTION

### 1.1. Background

Railway track as a base element of railway system greatly and directly influences safety and cost efficiency of rail transport. In process of track management, maintenance-of-way departments have to try to balance cost associated with potential damages arising from unfavorable tracks and cost for Maintenance & Renewal activities to minimize life cycle cost of track. To attain minimization of life cycle cost, there are key issues which need to be addressed. One of them is the railway track condition forecast technology [1].

In order to forecast railway track condition it shall be defined first. There are different methods of defining track condition but most of track condition forecasting models use track geometry parameters.

In order to measure track conditions by using track geometry model, typically track is divided into several shorter sections and geometry statistics are performed to each of them. Geometry statistics are then summed up to give a measure of overall segment quality, which is commonly called Track Quality Indices (TQIs). Use of TQIs provides possibility to assess railway track performance indicators, to design interventions, and to compare track performances before and after intervention[2]. Methods of calculating TQI varies by country. In China, TQI is calculated as the sum of the standard deviation of 7 track geometry measurement. In United States, TQI is calculated as ratio of traced space curve length to track segment length.

In Europe, J synthetic coefficient is used as an indicator of track quality based on standard deviation in Polish Railways. In India, a formula, called track geometry index (TGI), has been developed by Indian Railways to represent quality of track. This model is based on standard deviation of different geometry parameters over a 200m segment [3]. This paper discusses different methods used to represent track quality and analyses three of them based on real geometric measurement data from Addis Ababa-Djibouti railway for the purpose of recommending one method to be used for predicting the future state of track in the aim of implementing predictive maintenance.

### 1.2. Problem statement

Railway transport is the most economical transport next to water transport especially for freight transportation. To maintain the economic benefit of railway transport it is important to make the running cost as low as possible. One of the major running costs includes infrastructure maintenance cost which takes the greater share of infrastructure maintenance costs. To achieve minimization of maintenance cost it is important to implement predictive (condition based) maintenance which needs prediction of future state of infrastructure. Predicting needs prediction models and to have better models the condition of the structure shall be defined in a better indicator. In case of track infrastructure there are different methods representing quality of track geometry this paper focuses in discussing and recommending method of track geometry quality aiming to use it for predicting future state of track infrastructure. Prediction will help for implementation of predictive (condition based) track infrastructure maintenance.

### 1.3. Research purpose and objective

The purpose of the study is discussing different method of track quality index and recommending better method of quality index based on real track geometry measurement on existing railway line. The main objective is recommending track quality index method which will be used for prediction of future track

condition and supporting the implementation of predictive track infrastructure maintenance.

## 2. Research methodology

The study method consists of explorative research and statistical quantitative data analysis, explorative approach is used in literature reviews and quantitative statistical data analysis is used for comparison and recommendations based on results.

### 2.1. Research process

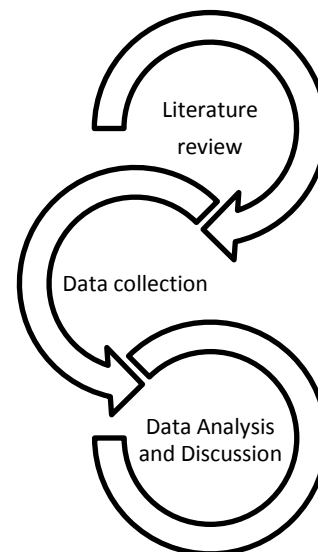


Figure 1 Research process flow chart

### 2.2. RESEARCH METHODS

Exploratory method used in literature review and statistical quantitative data analysis is used in data analysis. Literature study and survey is employed, the materials used include Journals, Thesis, Books, Manuals, conference papers. The lists of Key words used to search literatures:-*Track quality Index, Track geometry defects, track degradation, track quality level*

Secondary quantitative data is collected form Addis Ababa Djibouti railway line regarding the geometry of railway track from Track geometry measuring vehicle. Data analysis activities include data selection based on the line characteristics and classification by a section of 200m length, preliminary data analysis and detailed data analysis.

### 3. Railway track characteristics and degradation parameters

Changes in TQI, track settlement and average growth of track's irregularity are considered to be main track deterioration criteria from the aspects of track geometry, on tracks sub-structure and super structure, respectively[4].

Track geometry degradation is usually quantified by five track defects: the *longitudinal leveling defects*, the *horizontal alignment defects*, the *cant defects*, the *gauge deviations* and the *track twist*[5].

By looking to literature it can be observed that most of researchers considered short wavelength longitudinal level as crucial factor in degradation modeling[6]. This can be seen in

Figure 2

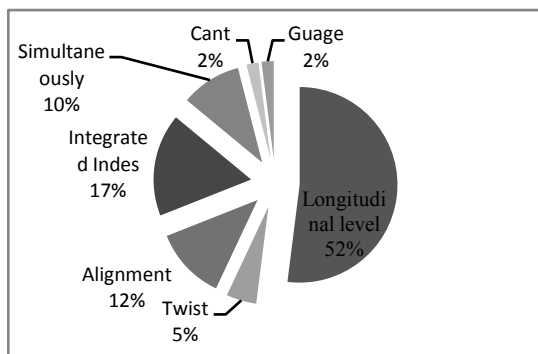


Figure 2 Distribution of applied track geometry measures [6]

#### 3.1. LITERATURE REVIEW

Determining an indicator to represent track quality is an essential prerequisite for modeling track degradation. Indices for representing track quality condition are demonstrated in Figure 3[6].

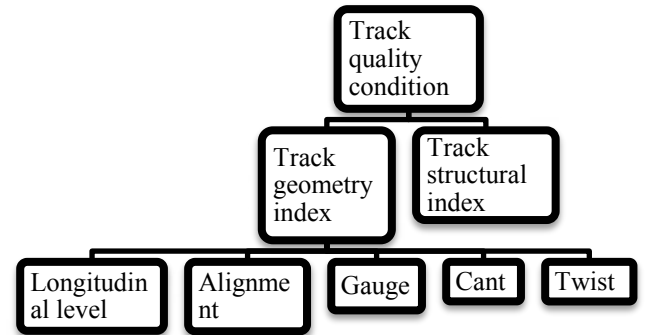


Figure 3 Track condition [6]

According to Xu [1], track condition is described by eight geometrical parameters : Gauge, Cross Level, Left/Right Surface, Left/Right Alignment, Twist, and Curvature[1]. But in most cases, an artificial track quality index (TQI) has been created as a linear combination of geometry measurements to indicate track state. These have been used in a Markov model where TQI is calculated in a range of 0-100 based on unevenness, twist, alignment and gauge measurements [7]. Track Quality Index is defined as a numerical value that represents the relative condition of track surface geometries [2]. American Railway Engineering and Maintenance-of-Way Association (AREMA) defined TQI as a number, derived from a formula that characterizes measured data collected from a Track Geometry Measurement Vehicle (TGMV) over a segment of track. It summarizes relatively large quantity of discrete measurements generated by a TGMV to allow characterization of an entire track segment [8].

#### 3.2. Space curve method

On study by Sharma, track geometry data for each 30.48cm is first aggregated into 160.934m segment, and each segment is  $L_0$  in length. TQI is then calculated for each type of track geometry measurement individually using the following formula.

$$TQI = \left( \frac{L_s}{L_0} - 1 \right) \times 10^6 \quad (1)$$

where

TQI = track quality index;  $L_s$  = traced length of space curve (m);  $L_o$  = fixed 160.934m length of track segment

$$L_s = \sum_{i=1}^n \sqrt{(\Delta y_i)^2 + (\Delta x_i)^2}$$

$$= \sum_{i=1}^n \sqrt{(\Delta y_i)^2 + 0.0929} \quad (2)$$

where

$\Delta y_i$  = difference in two adjacent measurements (m.);  $\Delta x_i$  = sampling interval along the track (=0.3048m.);  $i$ =sequential number.

In the presence of separate track geometry data for left track and right track, as in the case of surface and cant, we always choose the measurement (error) with higher absolute value[3].

As illustrated in Figure 4, for a specified track segment length, the rougher the track surface, the longer the space curve will be when stretched into a straight line[9].

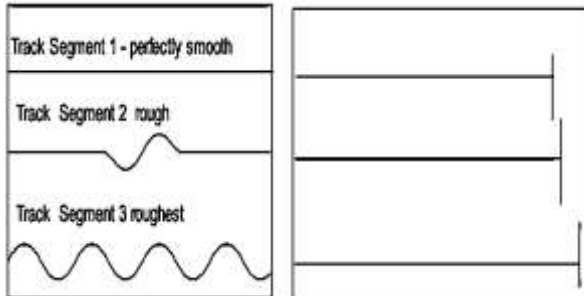


Figure 4 FRA Length-Base TQI Approach[9]

According to conclusions of this research “The TQIs developed were found to be able to quantitatively evaluate track quality and relate track quality to the Federal Track Safety Standards. These TQIs may be used to further evaluate vehicle and track interaction by incorporating vehicle characteristics. They may also be used as a tool to evaluate the effectiveness of track maintenance activities” [9].

### 3.3. J Synthetic Coefficient

In Europe, J synthetic coefficient is used as an indicator of track quality based on standard deviation in Polish Railways [3]. Four track

geometry parameters are considered in this index: vertical irregularities, horizontal irregularities, twist, and gauge [2]. The equation for calculating J synthetic coefficient is:

$$J = \frac{S_z + S_y + S_w + 0.5 * S_e}{3.5} \quad (3)$$

where: -  $S_z$ ,  $S_y$ ,  $S_w$  and  $S_e$ , are standard deviation of vertical irregularities, horizontal irregularities, twist, and gauge, respectively. Standard deviation for each measured parameter is calculated by the following equation: [2]

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (4)$$

Based on the above equation,  $n$  is identified as number of signals registered on track being analyzed,  $x_i$  represents value of geometry parameters at point  $i$  and  $\bar{x}$  is the average value of measured signals. J synthetic track quality coefficient also specifies allowable deviation of J [2]

Table 1. Allowable deviation of J coefficient based on line speed[2]

Speed [km/h]	J Coeff. [mm]	Speed [km/h]	J Coeff. [mm]
80	7.0	150	2.3
90	6.2	160	2.0
100	5.5	170	1.7
110	4.9	180	1.6
120	4.0	190	1.5
130	3.5	200	1.4
140	2.8	220*	1.1

\*Calculated through extrapolation

### 3.4. Track Geometry Index (TGI)

Indian Railways developed a formula to represent quality of track called TGI. This model is based on standard deviation of different geometry parameters over a stretch of 200 m segment. TGI is calculated for each segment and average value of such segments in every km gives general TGI value. With respect to effect of each geometry parameter on ride

quality, TGI has given different values for various geometry parameters as shown in the following formula:[2]

$$TGI = \frac{2UI + TI + GI + 6AI}{10} \quad (5)$$

where: - UI, TI, GI, and AI are index for unevenness, twist, gauge, and alignment respectively. For each measured track parameters, the index is calculated from the relation:

$$GI, TI, AI, UI = 100 \times e^{-\left(\frac{SD_{meas} - SD_n}{SD_{maint} - SD_n}\right)} \quad (6)$$

where:-  $SD_{meas}$  is standard deviation of measured geometry parameters,  $SD_n$  represents standard deviation prescribed for newly laid track and  $SD_{maint}$  is prescribed standard deviation for maintenance.  $SD_n$  and  $SD_{maint}$  are given in table 2

Table 2 Standard deviation (SD) values[2]

Parameters	Chord Length	SD for newly laid track	$SD_{maint}$	
			$V_{max} \geq 105$ km/h	$V_{max} < 105$ km/h
Unevenness	9.60	2.50	6.2	7.2
Twist	3.60	1.75	3.8	4.2
Gauge	1.00	1.00	3.6	3.6
Alignment	7.20	1.50	3.0	3.0

Table 3 TGI Classification for maintenance [2]

No	TGI Value	Maintenance requirement
1	$TGI > 80$	No maintenance required
2	$50 < TGI < 80$	Need basic maintenance
3	$36 < TGI < 50$	Planned Maintenance
4	$TGI < 36$	Urgent Maintenance

The advantages of TGI are:

1. It gives an idea of health of continuous length rather than highlighting isolated bad locations.
2. It gives due weightage to different parameters as per their effect on the Ride Index.
3. The range over which it varies is much smaller and it does not get affected by minor changes from run to run. A variation of 10 in

TGI shows a significant improvement/deterioration in the track quality [10].

### 3.5. Italian Railway Quality Indices

In order to calculate Rail quality indices RQI (Italian IQB), the Italian railway regulations on rail maintenance specify the following defectiveness indexes: *defectiveness index of longitudinal level*, equal to standard deviation on a 200m plane of longitudinal level; *defectiveness index of alignment*, equal to standard deviation on a 200m plane of alignment; *defectiveness index of transversal level*, equal to standard deviation on a 200m plane of transversal level; *wedging index*, equal to highest on a 200m plane, and therefore to the worst of the above-mentioned defectiveness indexes [11].

The regulations on rail maintenance survey, introduced by Italian Railway Network (*RFI, Rete Ferroviaria Italiana*), impose three **Rail Quality Levels** which call for “full implementation of the line” and a level which requires such railway operational restrictions as slowing downs on the line and traffic blocks [11].

Table 4 Levels of Degradation for Rail quality Index RQI (Italian IQB) [11]

Degradation level	Threshold value	Required action
Optimal Level	1.2	excellent geometry conditions
Level of attention	1.8	geometry is to be monitored
Level of intervention	2.25	maintenance works is required
Level of safety	2.7	traffic slowing down or block required

### 3.6. Five Parameters of Defectiveness

Five parameters of defectiveness are noted as  $W_5$ , which is a quality measure of line segments developed by Polish Railways. The formula treats defectiveness of each geometry parameter as an independent event in practice [12]. Considering arrangement of parameters

$$W_5 = 1 - (1 - W_e) \cdot (1 - W_g) \times (1 - W_w) \cdot (1 - W_x) \cdot (1 - W_y) \quad (7)$$

where: -  $W_e$  – defectiveness of track gauge,  $W_g$ – defectiveness of cant,  $W_w$ – defectiveness of twist,  $W_x$  and  $W_y$  are arithmetic averages for vertical and horizontal irregularities, respectively, as determined from defectiveness of left and right rails. Coefficient of parameter defectiveness  $W$  in the approach is calculated using the following Eq.8 [12]

$$W = \frac{\sum_{i=1}^n l_i}{l} \quad (8)$$

where: -  $W$  has to be substituted with  $W_e$ ,  $W_g$ ,  $W_w$ ,  $W_x$  and  $W_y$  respectively;  $l_i$  is a number of samples of assessment section which exceeded an allowed value of  $W_e$ ,  $W_g$ ,  $W_w$ ,  $W_x$  or  $W_y$  respectively;  $l$  is a total number of section samples,  $n$  is a number of exceedances of allowed threshold for total measured section. Five-parameter defectiveness is calculated based on the exceedances of the maximum allowed limit values. The qualification for line maintenance which depends on defectiveness value is specified in Table 5 [12].

Table 5 Quality qualification of track lines [12]

Evaluation of line	New	Good Condition	Sufficient condition	Indicating insufficient condition
Value $W_5$	$W_5 < 0.1$	$W_5 < 0.2$	$W_5 < 0.6$	$W_5 > 0.6$

### 3.7. TQI Proposed by A. Chudzikiewicz et al

A paper by Andrzej [12]. developed a new method of determining TQI by conducting a complete dynamic analysis of railway vehicle/track system response. In this system, defect and degradation of track are estimated from vertical acceleration measured on an axle-box. Algorithm proposed in the paper specifies TQI as determined by inertial measurement. Inertial measurement is based on a simple law where double integration of acceleration indicates a position on an accelerometer. For example, a vertical position of a wheel can be computed by double integration of axle-box acceleration. The result

provides longitudinal level due to a wheel being continuously in contact with a rail. TQI dependent on velocity takes form: -

$$TQI = W_{t,v}(v) = c_t \cdot \left( \pi \cdot \left( \frac{v}{v_e} \right)^6 \cdot \lim_{T \rightarrow +\infty} \left( \frac{1}{T} \int_0^T \dot{a}_e^2 \left( \frac{v}{v_e} \cdot t \right) dt \right) \right)^{0.15} \quad (9)$$

where:  $v_e$ –chosen reference velocity,  $v$  –current vehicle velocity,  $c_t$ –is constant value set on the basis of numerical research,  $t$  –time and  $a$ –axle box acceleration [12].

Table 6 Track quality qualification railway tracks by TQI coefficient [12]

Evaluation of line	New	Good Condition	Sufficient condition	Indicating insufficient condition
TQI Value	$< 0.1$	$< 0.13$	$< 2.20$	$> 2.20$

### 3.8. Australian Rail Track Corporation TQI

Australian Rail Track Corporation (ARTC) uses a ‘Track quality index’ (TQI) to provide an indication of track condition for specific sections of track. A TQI is derived from statistical analysis of track geometry car data for vertical alignment, horizontal alignment, twist and gauge over 100 m sections of track. Summation of four calculated indices provides a combined TQI for each 100 m section of track. Values are then averaged to give a TQI for longer sections of track or a rail corridor[13].

The intent of the TQI is not to provide a quantifiable pass/fail indication of track condition, nor is it used to identify specific track defects. TQI provides an overview of track quality and longer term trend analysis for strategic programming of track improvement works on the rail corridor. The ARTC typically reports on the percentage of track for each corridor that exceeds a TQI value of 25, considered (based on historical experience) as an optimal target maintenance level for concrete sleeper track. Specific track irregularities are identified through track

inspection and track geometry car exception reports[13].

### 3.9. Track quality Index used in China railways

TQI is the summation of standard deviations of seven irregularities, that is vertical irregularities (left and right) alignment irregularities (left and right), gauge, cross-level irregularity, and warp, in each 200m long track section[14]

$$TQI = \sum_{i=1}^7 \sigma_i \quad (10)$$

$$\sigma_i = \sqrt{\frac{1}{n} \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2} \quad \text{and} \quad \bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij} \quad (11)$$

where N (N=7) denotes irregularity number,  $\sigma_i$  is standard deviation of each irregularity,  $x_{ij}$  is the value of local irregularities,  $\bar{x}_i$  is average value deviation,  $n$  is the number of sampling points.

Table 7 Track Quality Index Management Value [15]

Speed class	Left high low mm	Right high low mm	Left orbit mm	Right orbit mm	Gauge mm	Level mm	Triangle pit mm	TQI Value
$V \leq 80$ km/h	2.2~2.5	2.2~2.5	1.8~2.2	1.8~2.2	1.4~1.6	1.7~1.9	1.9~2.1	13~15
$80 \text{ km/h} < V_{\max} \leq 120 \text{ km/h}$	1.8~2.2	1.8~2.2	1.4~1.9	1.4~1.9	1.3~1.4	1.6~1.7	1.7~1.9	11~13
$120 \text{ km/h} < V_{\max} \leq 160 \text{ km/h}$	1.5~1.8	1.5~1.8	1.1~1.4	1.1~1.4	1.1~1.3	1.3~1.6	1.4~1.7	9~11

160km/h $V_{\max}$	1.1~1.5	1.1~1.5	0.9~1.1	0.9~1.1	0.9~1.1	1.1~1.3	1~1.4	7~9
-----------------------	---------	---------	---------	---------	---------	---------	-------	-----

Note: - the reference is translated from Chinese version to English by Google translate

### 3.10. Track quality Number MDZ

The MDZ number comprises both horizontal and vertical deviations in track together with speed and lack of super elevation. This measurement is developed to capture changes in acceleration over a certain distance from a passenger point of view by direct mathematical analysis of real track geometry data, recorded by measuring wagon. The variation of acceleration is regarded as main criteria for comfort. Therefore, the sum of all changes in acceleration over a certain distance (charged with some corrective parameters) reflects the MDZ number for this section. This quality number reflects the riding comfort [16]. The MDZ number is defined as

$$MDZ = c \times \frac{1}{L} \times v^{0.65} \times \sum_{i=1}^{\frac{L}{\Delta x}} \sqrt{(\Delta \text{vert. level})^2 + (\Delta \text{horiz. level} + \Delta \text{cant})^2} \quad (12)$$

where: -  $\Delta \text{vert. level}$  and  $\Delta \text{horiz. level}$ , is the difference in track deviation from one measurement point to the next. Here,  $\Delta \text{cant}$  is the difference in cant level from one measurement point to the next.

### 3.11. Q Index

Pro rail of Netherlands converts SD index into a more universal form across different classes of tracks, as shown in (13). Q index ranges from 10 to 0. The larger the Q index, the better track quality [17].

$$N = 10 * 0.675^{\frac{\sigma_i}{\sigma_i^{80}}} \quad (13)$$

where: -  $N$  denotes Q index for quality parameter over 200m track segment,  $\sigma_i$  is standard deviation for the quality parameter, and

$\sigma_i^{80}$  represents 80<sup>th</sup> percentile of standard deviations for 200m segments in maintenance section ranging from 5 to 10 km.

### 3.12. Canadian National Railway's TQI

Canadian National Railway Company uses 2<sup>nd</sup> order polynomial equation of standard deviation  $\sigma_i$  of measurement values for quality parameter over track segment to assess its partial quality, as formulated in (14). The overall quality assessment is achieved by averaging six partial quality indices for gauge, cross level, left (right) surface, and left (right) alignment[17].

$$TQI_i = 1000 - C * \sigma_i^2 \quad (14)$$

where  $C$  is a constant and takes value of 700 for main line tracks

A larger track quality index implies track segment has better quality.

## 4. Analysis of TQI measurement methods

In this section, three methods of TQI measurements are analyzed based on real track geometry measurement data from conventional railway line. The data is taken from Track geometry measuring vehicle record of Addis Ababa Djibouti standard gauge railway line on May 25-27, 2020. The section of reading on the line is from Adama (KM 114) to Mieso (Km 329+057)

The methods analyzed are J synthetic coefficient, track geometry index (TGI), and TQI used in Chinese railway.

### 4.1. Study area for the case comparison

The study area used for data collection is in Ethiopia and starts from Sebeta, 10Km distance from Addis Ababa passes through Debrezeyt, Mojo dry port, Addama, Metehara, Awash, Mieso, Dire Dawa and Dewale cities.



Figure 5 study area (Addis -Djibouti Railway line) source ERC website

Data collected from Addis Ababa – Djibouti railway line under operation is used for analysis. This line is about 656km in length and it is Chinese class II standard gauge railway. Track inspection vehicle checks dynamic partial unevenness (peak management) of track; involving track gauge, level, height, track-alignment, twist, vertical acceleration and horizontal acceleration.

The dynamic quality of overall unevenness (mean value management) at line section is assessed through track quality index (TQI). From these detailed data this study focuses on five of the track geometry data only 1) longitudinal profile: vertical unevenness 2) Horizontal alignment 3) Gauge 4) Cant (supper elevation) and 5) twist: the difference between supper elevation of the rail in two consecutive measurements (change in supper elevation) The track geometry measuring vehicle collects four of this basic data namely longitudinal profile, horizontal profile, Gauge, and cant others are derived from these basic data. From the whole line data used for the comparison of the selected three methods of track quality index is the section from Adama (KM 114) to Mieso (Km 329+057) which is shown by the rectangular box on figure 5. This covers about 215Km length.



## RESULTS

Table 8 Comparison between Chinese TQI, J synthetic value and TGI

Mileage (KM)	TQI		J synthetic		TGI	
	Value	Exceed standard	Value	Exceed standard	Value	Exceed standard
114	15.75	More than 10%	3.884	Not Exceed	54.27	basic maintenance
114.2	16.42	More than 10%	3.966	Not Exceed	52.86	basic maintenance
248.2	11.86	Not exceeded	3.04	Not Exceed	66.9	basic maintenance
300	14.64	exceeded	3.494	Not Exceed	64.53	basic maintenance
320.4	15.48	More than 10%	3.493	Not Exceed	79.88	basic maintenance
323.2	19.88	More than 20%	4.901	Not Exceed	45.48	Planned maintenance
324	16.21	More than 10%	3.854	Not Exceed	64.81	basic maintenance
324.2	20.34	More than 20%	5.061	Not Exceed	43.74	Planned maintenance

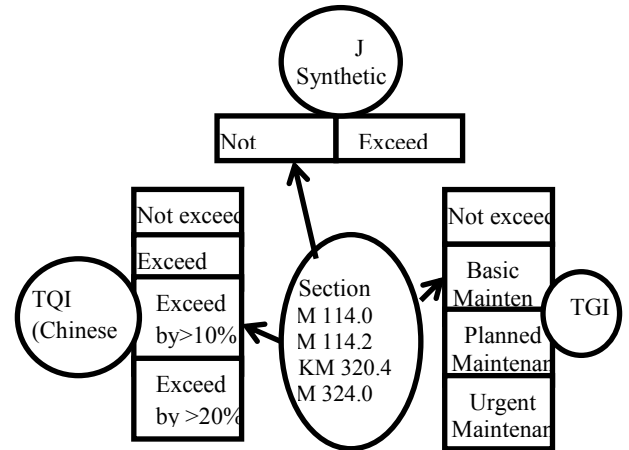


Figure 6, Comparison of TQI, TGI, and J Synthetic methods by severity level, case 1

As we can see from figure 6 in this situation the Chinese TQI measurement method seems more conservative than the other two methods this is because TGI method gives wider range for the “Basic maintenance” level of severity. J synthetic method only reflects two categories of track quality one above the threshold and the other below the threshold value which makes the categorization more difficult in understanding the progress of deterioration in several levels of quality.

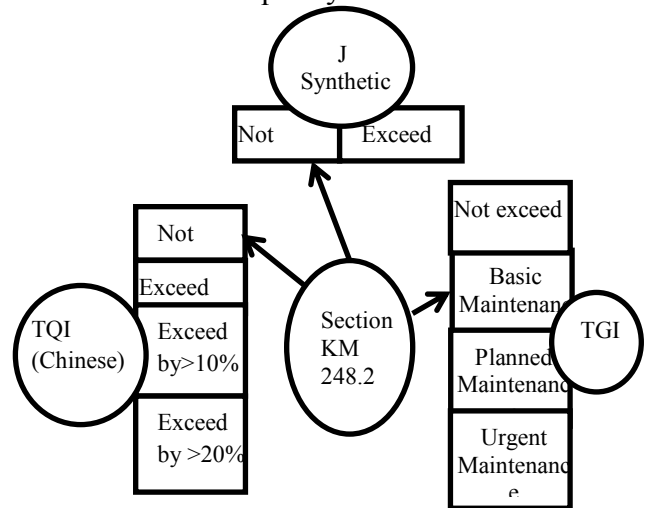


Figure 7 Comparison of Chinese TQI, TGI and J Synthetic methods case 2

On this case the reverse of case one above becomes evident. TGI categorizes this section in to more Sevier stage than any of the two methods, this happens because TGI gives more weightage to track geometry parameters which have more serious effect on ride quality and

less to those having less effect on ride quality during the calculation of TGI value.

However the Chinese TQI method gives equal weight for all parameters of track geometry degradation. We can say that TGI method is safer for this case.

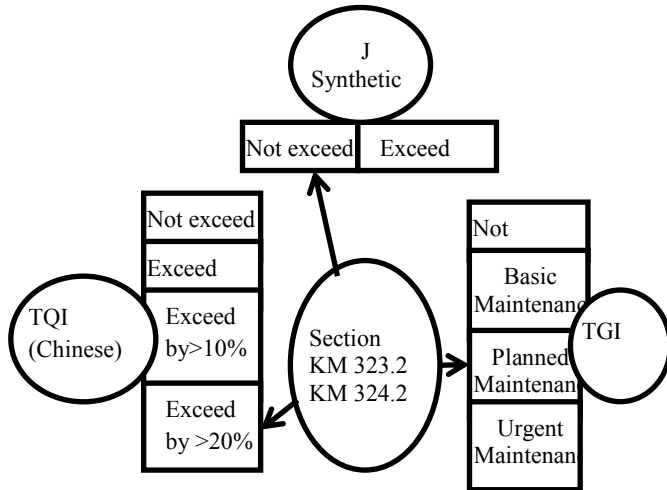


Figure 8 Comparison of TGI, Chinese TQI and J Synthetic methods case 3

In this case, similar to the case one above, TQI seems more conservative than the other two methods. This happens also because of giving the same weightage for all parameters of degradation during the calculation process of the aggregated TQI value.

## DISCUSSIONS

Because the line in consideration for the case study is a conventional track and the design speed for freight is 80Km/hr. A speed of 80Km/hr value is taken for the limits of the track quality index values in three of the methods under consideration. Even if the data used for analysis is taken on 215Km length of track, only some important sections are presented on the result.

As we can see from Table 8, J synthetic coefficient method is not effective in indicating detailed values of track quality. It can only show those two levels of track quality, one exceeding threshold value and the other not exceeding, because it has only one threshold value. Because of this J synthetic method is indicating safe values for all of the seven track

quality values which are above the acceptable limit in both of the other methods.

On the other hand most of the track sections considered out of the threshold value by the Chinese TQI method is also above the threshold by TGI (track geometry index) method too. Except one point at KM 248.2 at this section the Chinese TQI method shows the section is safe for operation and does not exceed the threshold value but TGI method shows that the section needs basic maintenance. This result shows that TGI is more conservative and safer to use than the two other methods of track quality index.

Regarding maintenance priority and definition of damage severity, both the TGI and Chinese TQI give different interpretations. For example if we take the section at KM 323.2 and KM 324.2 graphically represented on figure 8, the Chinese TQI method defined the damage as more serious by labeling the damage exceeded by more than 20% of the threshold level, which implies; this sections need more priority for maintenance but TGI method labels these two sections in medium level of severity, labeling them as they can be considered in planned maintenance.

These different interpretations come from the difference of calculation methods. The Chinese TQI method gives equal weights for all of the defects i.e. horizontal unevenness (right and left), Vertical unevenness (right and left), level, gauge and twist. But TGI gives different weight for those geometric defects by allocating higher weight for more serious effects on ride quality and lower weight for those having less effect on ride quality. In these regards TGI is more reliable than the Chinese TQI method.

To the contrary on the section at KM 114, KM 114.2, KM 300, KM 320.4, and KM 324 the Chinese TQI method labels the defects as medium Severity defects by indicating “*more than 10% exceeded*” from the threshold, for all sections except KM 300 which is labeled as lower level severity damage. But TGI labels the section in “*Basic maintenance*” which indicates lower level of severity. In this regard the

Chinese TQI method seems more conservative than TGI. The reason behind is again on the calculation method of the aggregated TQI in which TGI gives more reasonable weightage for track geometry parameters rather than giving same coefficient for all.

## CONCLUSIONS

From the above results and discussion it can be concluded that both TGI and the Chinese TQI can give more categorization of track quality than J Synthetic method. But TGI gives a more reasonable track quality value than all of the three methods analyzed.

Ethio-Djibouti railway line uses Chinese TQI method for characterizing track geometry quality index which gives more categorization of track quality but it's advisable to include TGI method.

## LIMITATIONS OF THE STUDY

This paper analyzed three methods of track quality measurement even if it discussed more than ten methods of track quality measurement this is because most of the methods reviewed need a more advanced track geometry data and the data which can't be found easily and needs a more advanced track geometry measuring vehicle or special equipment for data collection. Hence it is recommend other researchers to compare more options of track geometry measurement methods.

## REFERENCES

- [1] Peng Xu, R.-K.L., Feng Wang, Fu-TianWang, and Quan-Xin Sun, *Railroad Track Deterioration Characteristics Based Track Measurement Data Mining*. Mathematical Problems in Engineering, 2013. **2013**(970573): p. 7.
- [2] Abdur Rohim Boy Berawi, R.D., Rui Calçada, Cecilia Vale, *Evaluating Track Geometrical Quality Through Different Methodologies*. International Journal of Technology, 2010. **1**: p. 11.
- [3] Sharma, S., et al., *Data-driven optimization of railway maintenance for track geometry*. Transportation Research Part C: Emerging Technologies, 2018. **90**: p. 34-58.
- [4] Askarinejad, J.S.a.H., *Influences Of Track Structure, Geometry And Traffic Parameters On Railway Deterioration*. IJE Transactions B, 2007. **20**(3): p. 10.
- [5] Andrade, A.R., *A Bayesian model for rail track geometry degradation: a decisive step towards the assessment of uncertainty in rail track life-cycle.*, in *12th WCTR*. 2010, Technical University of Lisbon: Lisbon, portugal. p. 20.
- [6] Ahmadi, I.S.a.A., *Current Trends in Reliability, Availability, Maintainability and Safety*. 2016 ed. A Survey on Track Geometry Degradation Modelling. 2016, Switzerland: Springer International Publishing. 10.
- [7] John Andrews, D.P.a.F.D.R., *A Stochastic Model for Railway Track Asset Management*. Reliability Engineering and System Safety, 2014. **130**(0951-8320): p. 9.
- [8] AREMA, *Manual for Railway Engineering, in Systems Management*. 2010, American Railway Engineering and Maintenance-of-Way Association: USA. p. 736.
- [9] Administration, F.R., *Development of Objective Track Quality Indices*. 2005, US Department of transportation Federal Railroad Administration: USA. p. 5.
- [10] Mundrey, J.S., *Railway Track Engineering* 4ed. 2010, New Delhi: Tata McGraw Hill Education Private Limited. 672.
- [11] Ferdinando Corriere, D.D.V., *The Rail Quality Index as an Indicator of the "Global Comfort" in Optimizing Safety, Quality and Efficiency in Railway Rails.*, in *SIIV - 5th International Congress - Sustainability of Road Infrastructures*,

- SIIV2012 and S. Committee, Editors. 2012, Elsevier Ltd.: Palermo 90100, Italy. p. 10.
- [12] Andrzej Chudzikiewicz, R.B., Mariusz Kostrzewski, Robert Konowrocki, *Condition Monitoring Of Railway Track Systems*. Transport 2018. **33**(2): p. 12.
- [13] Bureau, A.T.S., *Safety of rail operations on the interstate rail line between Melbourne and Sydney*, A.T.S. Bureau, Editor. 2013, Australian Transport Safety Bureau: Australia. p. 104.
- [14] Limei Guo, H.L., Xianghua Wu, Hanyu Cui, *Study on Comprehensive Evaluation Method for Track Irregularity Based on HSMM*, in *4th International Conference on Sensors, Measurement and Intelligent Materials (ICSMIM 2015)*. 2015, The authors - Published by Atlantis Press: china. p. 4.
- [15] Corporation, M.o.I.a.E.o.C.R., *General speed railway line repair rules*, in *General speed railway line repair rules*. 2019, Ministry of Industry and Electricity of China Railway Corporation Beijing,. p. 143.
- [16] Lyngby, N., *Railway Track Degradation: Shape and Influencing Factors*. International Journal of Performability Engineering, 2007. **5**(2): p. 10.
- [17] Reng-Kui Liu, P.X., Zhuang-Zhi Sun, Ce Zou, and Quan-Xin Sun, *Establishment of Track Quality Index Standard Recommendations for Beijing Metro*. Discrete Dynamics in Nature and Society, 2015. **2015**: p. 9.

# ANALYTICAL STUDY ON SEISMIC PERFORMANCE OF PARTIALLY PRESTRESSED CONCRETE BEAM-COLUMN JOINTS

Hilina Assega<sup>1</sup> and Adil Zekeria<sup>1</sup>

<sup>1</sup> School of Civil and Environmental Engineering, Addis Ababa Institute of Technology,  
Addis Ababa University, Addis Ababa, Ethiopia

Corresponding author's email assega.hilina@gmail.com

## ABSTRACT

In this study, six partially prestressed concrete exterior Beam-Column Joints with variable prestressing force and four partially prestressed concrete interior beam-column joints with variable column axial load ratio have been analytically analyzed and assessed to evaluate their hysteretic performance under reversed cyclic loading. A Two-Dimensional finite element software program, VecTor2, is used to validate the non-linear response of beam-column joint experiments executed in Chiba and Kyoto university, Japan. The analytical result adequately simulated the interior joints in all cycles of loading while in the exterior joint a reasonable underestimation of results was obtained at the last cycle. In the partially prestressed concrete exterior beam-column joints, variation of prestressing force had little effect on the ultimate storey shear capacity. Stiffness and ductility increased significantly with prestressing force before wide shear crack formation and high prestress loss at the joint. Strength degradation after peak response was severe with increasing prestress level. This phenomenon undermined the inelastic energy dissipation capacity of the highly prestressed joints at the later cycles. The increment of column axial load in the partially prestressed concrete interior beam-column joint resulted in wider pinching while the converse created severe pinching. Premature crushing of concrete at the joint

due to high compressive stress from the column axial load and prestressing force was not observed in any of the specimens.

**Keywords:** Partially Prestressed Concrete, Beam-Column joint, Prestressing force, Column axial load ratio, VecTor2, hysteretic response.

## INTRODUCTION

For many years, the use of prestressed concrete members has been accepted to be advantageous for structures under flexure since it counteracts externally applied gravity loads. However, after a while, their significance in primary seismic-resistant members such as frames and shear walls has created a substantial argument. In Nishiyama's research [1], a partially prestressed concrete exterior beam-column joint, prestressed with  $0.5f_{yp}$  ( $f_{yp}$  is the yield strength of prestressing steel), was subjected to a reverse cyclic loading and showed improved hysteretic performance than ordinary reinforced concrete joints. However, these joints were designed to fail in flexure with plastic hinge occurring at the beam-column interface. Thus, the true shear behavior of the joint was not fully understood. Kashiwazaki and Noguchi [2] studied the effect of prestressing force level on the ultimate storey shear capacity of partially prestressed concrete interior beam-column joints and concluded that no significant effect was obtained.

According to Paulay and Priestley [3], many beam-column joint failures have been observed in the 1980 EI Asnam [4] earthquake. Shear and anchorage failures, particularly at exterior joints, have also been identified after the 1985 Mexico [5], the 1986 San Salvador [6], and the 1989 Loma Prieta [7] earthquakes. Beam-column joints are very critical regions in reinforced concrete frames designed for inelastic response to seismic excitation. This is because they are located in an area, where shear and bond stresses are considerably high. These forces are resisted by diagonal compression strut mechanism and truss mechanism. The diagonal compression strut resistance mechanism is fully active before the stress transfer mechanism at the joint is demolished due to degradation in bond strength. The performance of a diagonal compression strut mechanism can be improved by confining the joint. Usually, joint confinement is provided by adding more transverse reinforcements. In addition to that, as discussed in Park and Paulay [8], compressive stress from column axial load widens the diagonal strut region in the joint as a result of an enlarged compression block across the column region. Due to the formation of a wider diagonal compression strut, horizontal bond forces along the longitudinal beam bars can now be disposed of more easily.

In partially prestressed concrete beam-column joints, additional joint compressive stress is provided from the compression force due to post-tensioning. By considering this advantage, in this study, the effect of introducing compressive stress in the partially prestressed concrete exterior beam-column joint with variable prestressing force is studied and their hysteretic behavior is evaluated. Premature crushing of concrete in partially

prestressed concrete interior beam-column joint, that might occur due to high compressive stress, obtained from column axial load and prestressing force is also covered in this study.

## 1. FINITE ELEMENT MODELING AND VALIDATION

### Description of the examined partially prestressed concrete joints

Two experimental programs, partially prestressed concrete interior and exterior beam-column joint, executed by Kashiwazaki and Noguchi [2], and Nishiyama and Wei [9] in Chiba and Kyoto University, Japan, were used to validate the analytical result. Both joints are designed to fail in shear at the joint to evaluate the joint's ultimate capacity. The interior specimens were subjected to a constant column axial load of 320 kN while no column axial load was applied to the exterior specimens. Figure 1 and Figure 2 show the sectional detail and reinforcement layout of partially prestressed concrete interior and exterior beam-column joints. Table 1 and Table 2 shows the material properties of the partially prestressed concrete interior and exterior beam-column joints.

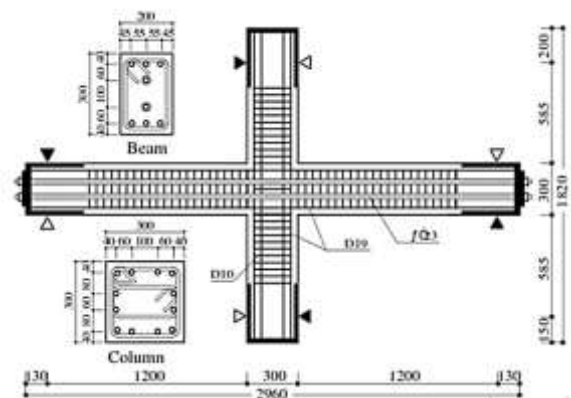


Figure 1 Sectional detail, reinforcement layout, and test setup of the partially prestressed concrete interior beam-column joint



# Analytical Study on Seismic Performance of Partially Prestressed Concrete Joints

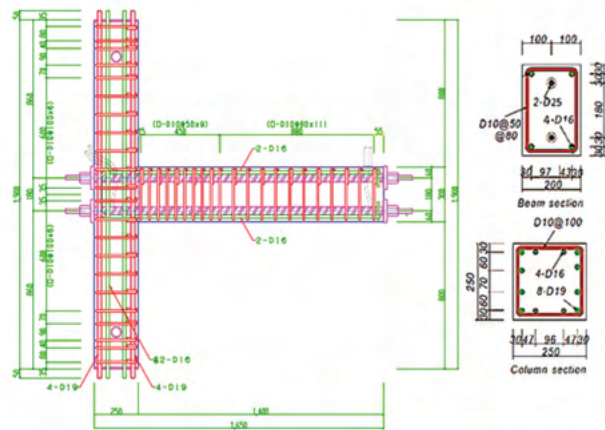


Figure 2 Sectional detail, and reinforcement layout for partially prestressed concrete exterior beam-column joint

Table 1 Material properties of partially prestressed concrete interior beam-column joint specimens

a) Concrete				
Specimen	Compressive strength, $f_c$ (MPa)	Elastic modulus, $E_c$ (GPa)	Split strength, $f_t$ (MPa)	Prestress (MPa)
PC-0	34.4	28	3.01	0
PC-1	34.4	28	3.01	$0.25f_{yp}$
PC-2	35.3	28.1	3.01	$0.5f_{yp}$

b) Reinforcement and Prestressing steel					
Type of reinforcement	Diameter (mm)	Yield strength, $f_y$ or $f_{yp}$ (MPa)	Yield strain, $\epsilon_y$ (mm/m)	Elastic modulus of steel, $E_s$ (GPa)	Ultimate tensile strength, $f_u$ (MPa)
Longitudinal	19	517	2.85	181	692
Transverse	10	897	4.33	207	1070
Prestressing steel	23	1100	5.5	200	1250

N.B:  $f_y$  is yield strength of reinforcement and  $f_{yp}$  is yield strength of prestressing steel

Table 2 Material properties of partially prestressed concrete exterior beam-column joint specimen

a) Concrete					
Specimen	Compressive Strength, $f_c$ (MPa)	Strain at $f_c$ , %	Elastic Modulus of concrete, $E_c$ (GPa)	Split tensile Strength, $f_t$ (MPa)	Prestress (MPa)
KPC2-1	34.6	0.22	28.2	2.51	$0.59f_{yp}$

b) Reinforcement and Prestressing steel					
Type of reinforcement	Diameter (mm)	Yield stress, $f_y$ / $f_{yp}$ (MPa)	Yield strain, $\epsilon_y$ (mm)	Elastic modulus of steel, $E_s$ (GPa)	Ultimate tensile strength, $f_u$ (MPa)
Transverse	10	307	1.744	176	436.9
Longitudinal	16	374.7	2.025	185	533.4
Longitudinal	19	386.8	2.113	183	569.9
Prestressing steel	25	1026	5.104	201	1146

N.B:  $f_y$  is yield strength of reinforcement and  $f_{yp}$  is yield strength of

## Material modeling

The developed finite element model sufficiently captured the material nonlinearity of concrete, reinforcement, prestressing steel, and the bond characteristic between the steel and the surrounding concrete by assigning the appropriate material models for each material property. A bond link element was assigned between the reinforcement/prestressing steel and the surrounding concrete, at the critical regions of the partially prestressed concrete exterior beam-column joint.

Their bond-slip behavior was defined by assigning the appropriate confinement pressure index factor according to the equation provided under F. J. Vecchio et al [10]. The reinforcements in the interior joints were assumed to be perfectly bonded since the confinement pressure provided by the transverse reinforcements exceed the high confinement pressure value (7.5 MPa) accepted by Committee Euro-International Du Beton, CEB-FIP Model Code 90 [11]. Concrete regions were modeled by using a rectangular concrete element and transverse reinforcements were defined as smeared elements inside the confined regions of the concrete by defining the appropriate transverse reinforcement ratio. Reinforcement and prestressing steel regions were modeled as truss bar elements.

### Mesh, constraints and loading

In the partially prestressed concrete interior beam-column joint a rectangular mesh type was used with an element size of 40mm by 40mm in all regions except the load-bearing region at the beam ends and top and bottom of the column, for which is 23x40mm and 25x40mm were used. Four load cases were defined. Load case 1 was a monotonic loading which was assigned for the constant column axial load and the own weight of the specimen. Load case 2, 3 and 4 was a reverse cyclic loading that represents the lateral storey displacement history,  $\pm 7, \pm 14, \pm 21, \pm 28, \pm 45$  and  $\pm 72$ mm, which was applied progressively at the beam ends. Moment release was provided at the end of the beams and at the top and bottom of the column to simulate the inflection points in the actual structure. A push and pull load was assigned at the tip of the beam to simulate load reversal due to alternative drift. Figure 3 shows loading and boundary conditions for this specimen.

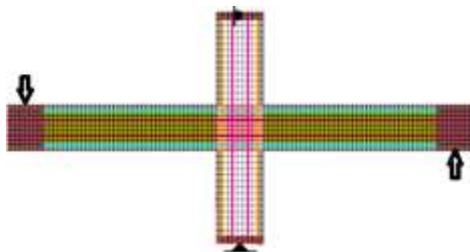


Figure 3 Loading and boundary conditions

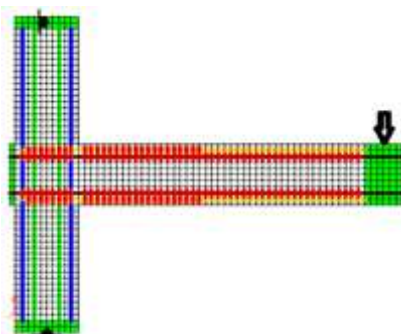


Figure 4 Bond link element, loading and boundary conditions

In the exterior joint a rectangular mesh was used with an element size of 25x30mm at the beam region and 30x30mm for the rest of the regions. Three load cases were defined. Load case 1 was a monotonic loading which was assigned for the self-weight of the specimen. Load case 2 and 3 was a reverse cyclic loading that represents the lateral storey displacement history,  $\pm 12, \pm 24, \pm 36, \pm 48, \pm 60$  and  $\pm 90$ mm, which was applied progressively at the beam ends. Figure 4 shows bond link element, loading, and boundary conditions for these specimens.

### Analytical and Experimental result comparison

In general, the finite element model sufficiently simulated the experimental result in both partially prestressed concrete interior and exterior specimens. The analytical hysteretic curves for the interior specimens (PC-0, PC-1, and PC-2) and exterior specimens (KPC2-1) in comparison with the experimental result are presented in Figure 5. The error that occurred in specimen KPC2-1 at the last cycles of the loading was significant relative to the others. This is due to the software's inadequate performance when the specimen's material nonlinearity becomes more pronounced as the number of repeated loading cycles increases after peak capacity. Such errors are believed to be improved by executing a three-dimensional analysis. Since two-dimensional models do not capture the effect of out-of-plane crack propagation, they tend to exaggerate surface crack width which might in turn underestimates the capacity of the member. The analytical failure crack pattern observed at each specimen in comparison with the experimental is also presented in Table 3.



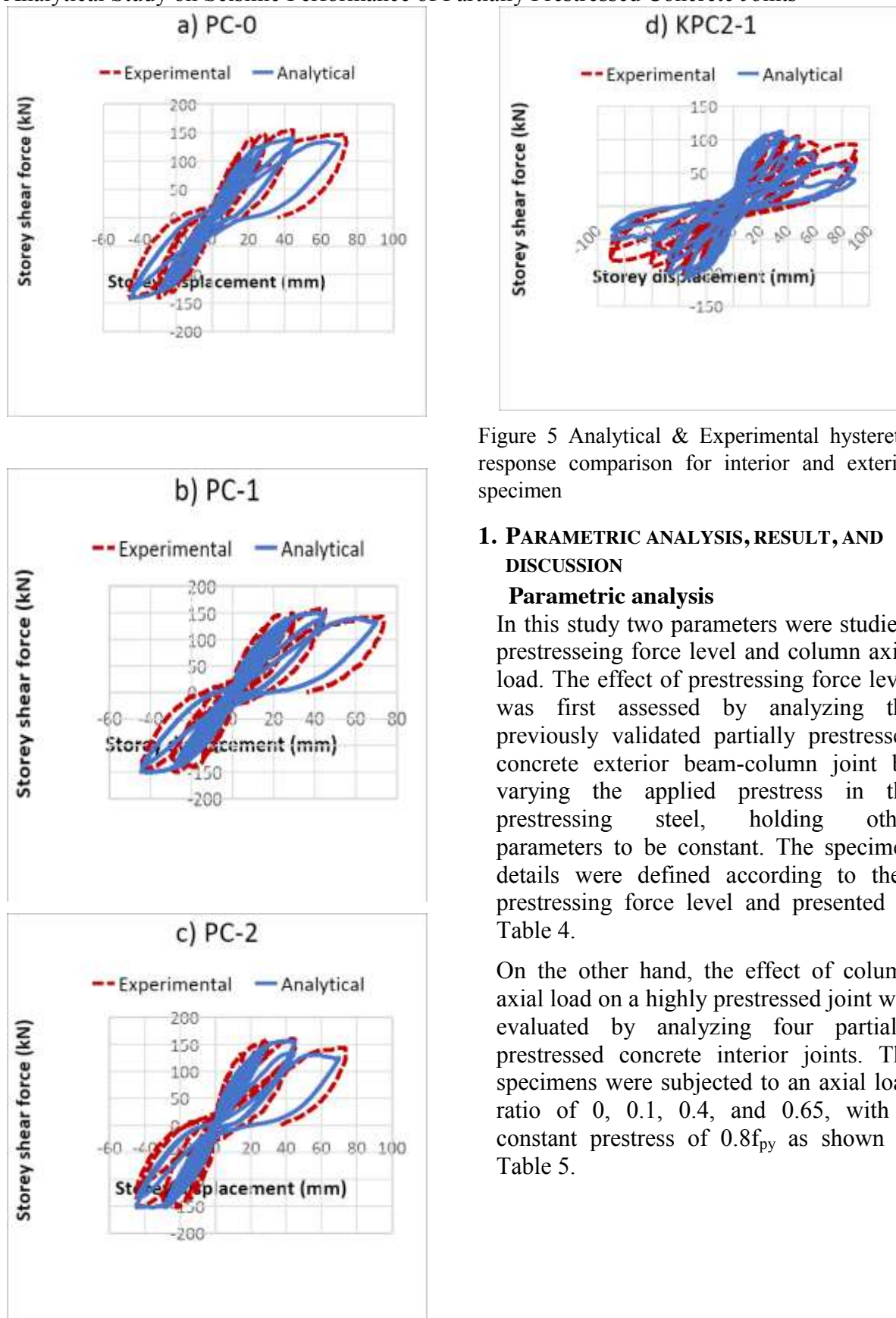


Figure 5 Analytical & Experimental hysteretic response comparison for interior and exterior specimen

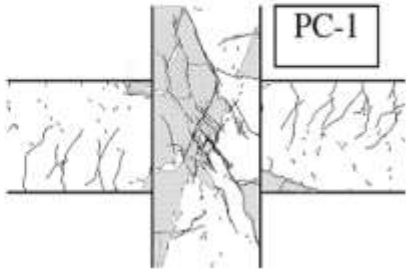
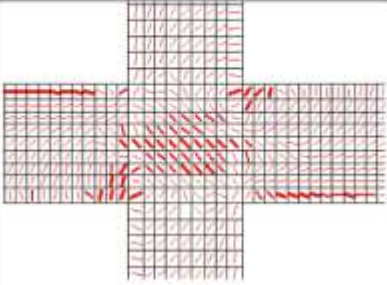
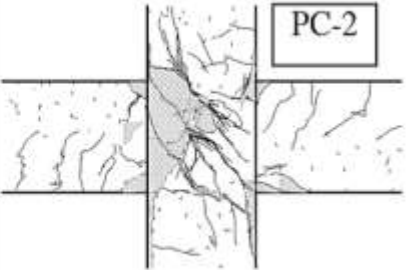
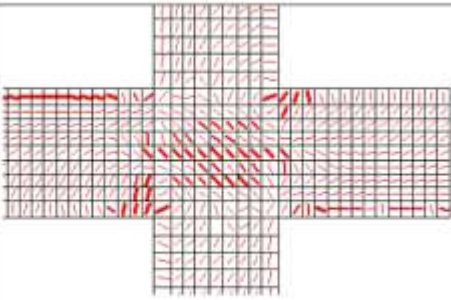

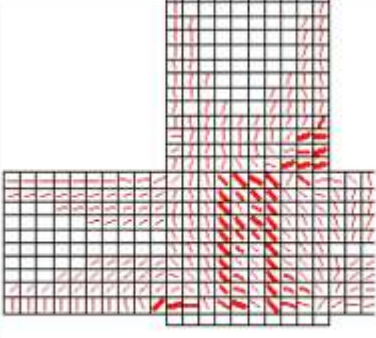

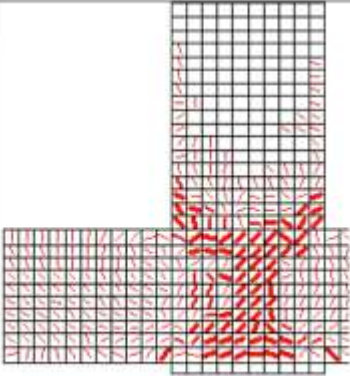
## 1. PARAMETRIC ANALYSIS, RESULT, AND DISCUSSION

### Parametric analysis

In this study two parameters were studied, prestressing force level and column axial load. The effect of prestressing force level was first assessed by analyzing the previously validated partially prestressed concrete exterior beam-column joint by varying the applied prestress in the prestressing steel, holding other parameters to be constant. The specimen details were defined according to their prestressing force level and presented in Table 4.

On the other hand, the effect of column axial load on a highly prestressed joint was evaluated by analyzing four partially prestressed concrete interior joints. The specimens were subjected to an axial load ratio of 0, 0.1, 0.4, and 0.65, with a constant prestress of  $0.8f_{py}$  as shown in Table 5.

Table 3 Failure crack pattern comparison for the experimental and analytical model

Specimen	Experimental	Analytical
PC-1 (at a storey displacement of 74mm)		
PC-2 (at a storey displacement of 74mm)		
KPC2-1 (at a storey displacement of 36mm)		
KPC2-1 (at a storey displacement of 90mm)		

### Analytical result

The seismic performance of the joints was evaluated based on the storey shear force - storey displacement hysteretic response. The storey displacement is the displacement equal to the lateral displacement history applied at the tip of the beam and the storey shear force is the lateral reaction at the tip of the beam under the applied lateral displacement. Figure 6 shows the storey shear force vs storey displacement hysteretic response of partially prestressed concrete exterior joints. Figure 7 shows the hysteretic response for partially prestressed concrete interior joints.

Table 4 Summary of prestress loading on partially prestressed concrete exterior joint

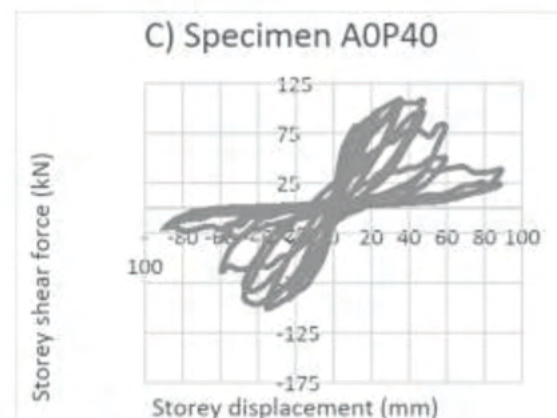
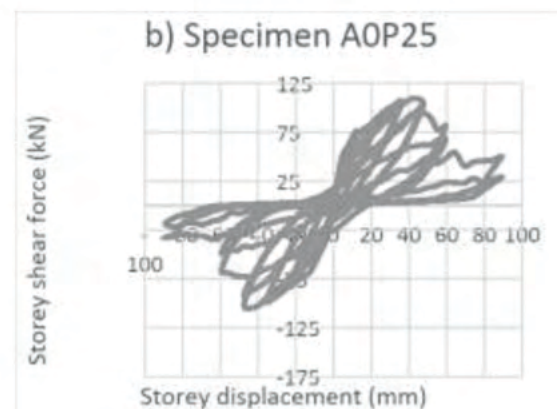
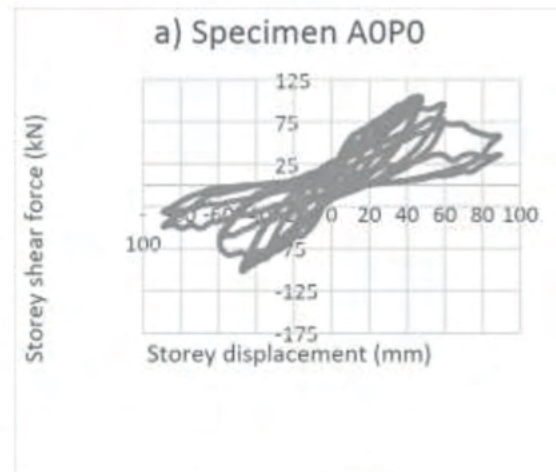
Specimen	A0P0	A0P25	A0P40	A0P59	A0P70	A0P80
Prestress ratio (%)	0	25	40	59	70	80
Prestress $\sigma_{ps}$ (MPa)	0	256.5	410.4	615.6	718.2	820.8
(% $f_{py}$ )						

N.B:  $f_{py}$  is Prestressing steel (PS) yield strength

Table 5 Summary of column axial load on partially prestressed concrete interior joint

Specimen	IA0P80	IA10P80	IA40P80	IA65P80
Column axial load, $N_{ED}$ (kN)	0	320	1238	2012
Axial load ratio, $N_{ED}/(A_c \cdot f_{cd})$	0	0.1	0.4	0.65

N.B:  $A_c$  is the gross sectional area of the column and  $f_{cd}$  is the design compressive strength of concrete





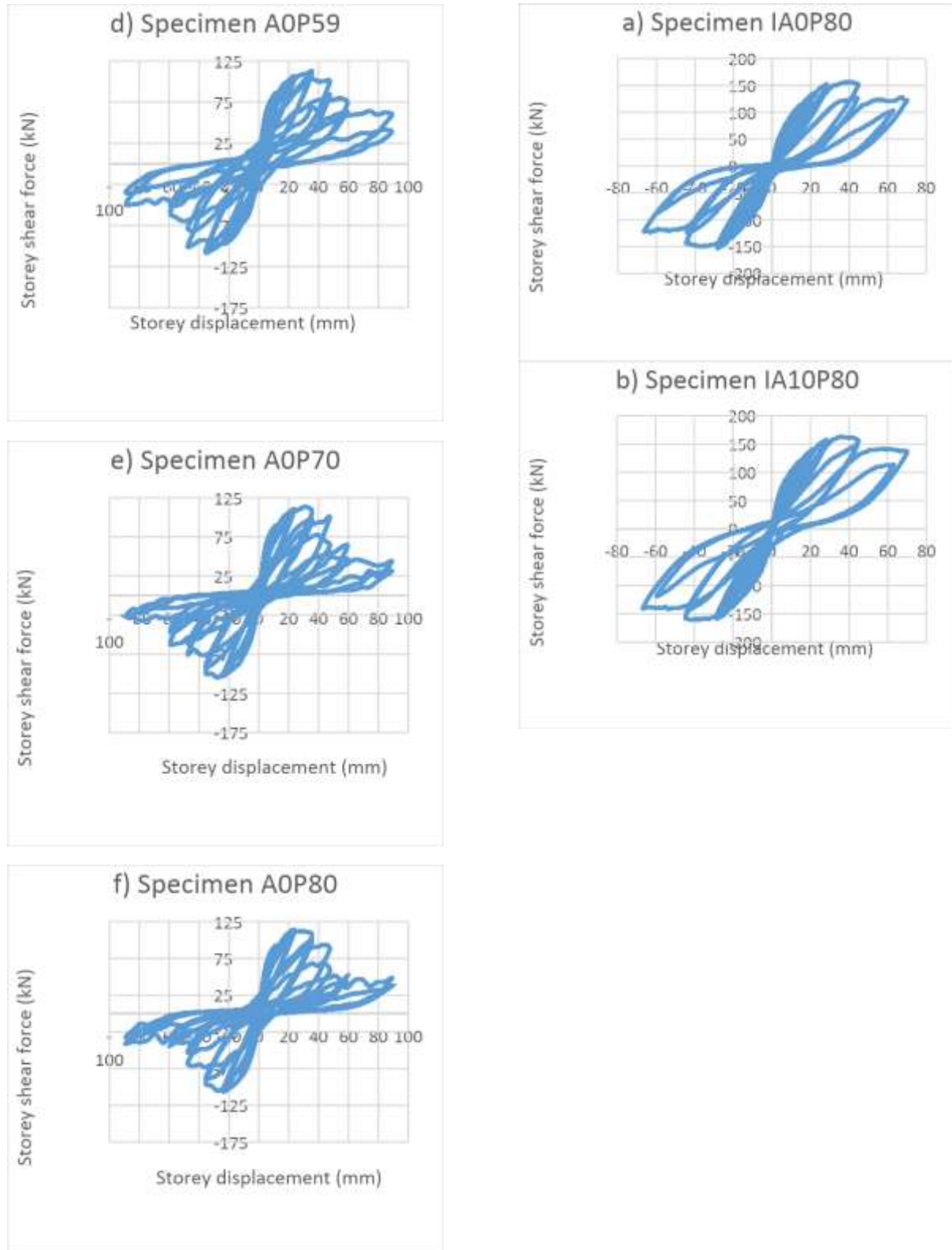


Figure 6 Hysteresis response of partially prestressed concrete Exterior beam-column joint

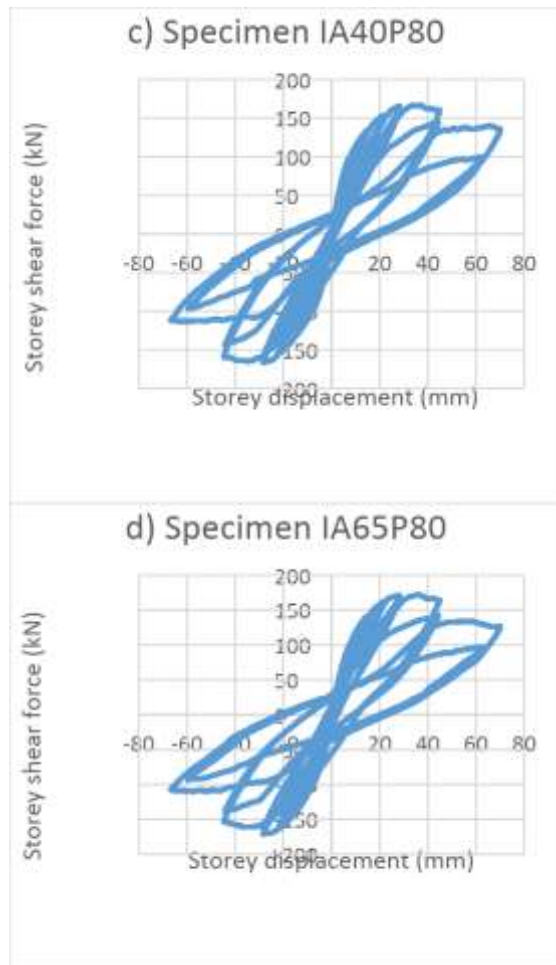


Figure 7 Hysteretic response of partially prestressed concrete Interior beam-column joint

## DISCUSSIONS

### I. Effect of prestressing force on the partially prestressed concrete Exterior beam-column joint

The seismic performance was evaluated based on storey shear capacity, stiffness, shear strength degradation after peak response, ductility, absorption, and energy dissipation capacity Ultimate storey shear capacity.

#### (a) Ultimate storey shear capacity

As can be observed from Figure 8, the prestressing force resulted in an insignificant effect on the ultimate storey shear capacity, under both positive and negative loading protocol. This is because

under higher cyclic displacement prestressed concrete sections resemble the behavior of reinforced concrete sections. Therefore, they will have more or less similar capacity after experiencing some cycles.

#### (b) Ductility, Stiffness, Strength degradation, and Energy dissipation capacity

High ductility is essential in seismic design to delay the local failure of members by allowing plastic redistribution of actions from one critical section to another and to allow absorption and dissipation capacity of the input energy.

Figure 9 shows the effect of prestressing force on the displacement ductility factor. It can be seen that the increase in prestressing force in the beam-column joint resulted in a substantial increment in the ductility of the specimen. This is mainly due to the significant variation in the yield displacement. Due to the initially applied compressive stress by the prestressing steels, highly prestressed joints were able to counteract the positive deformation from the external loads at the initial stages of the loading. This led joints with higher prestress level to achieve their yield and peak/ultimate capacity at smaller yield and peak/ultimate deformations ( $\Delta_y$  and  $\Delta_{peak}$ ) as shown in Table 6.

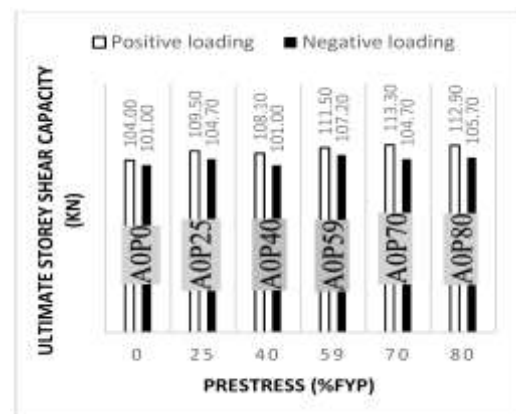


Figure 8 Effect of prestressing force on the ultimate storey shear capacity of partially

prestressed concrete Exterior beam-column joint

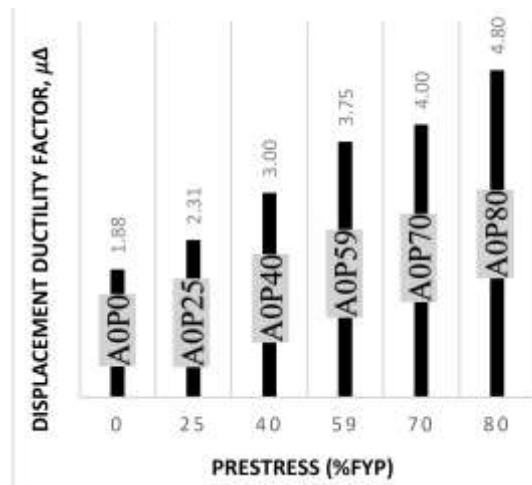


Figure 9 Effect of prestressing force on displacement ductility factor

Table 6 Yield and Ultimate displacements of partially prestressed concrete exterior joint

Specimen	A0P0	A0P25	A0P40	A0P59	A0P70	A0P80
$\Delta_y$ , (mm)	32	26	20	16	12	10
$\Delta_{@75\%Peak}$ , (mm)	60	60	60	60	48	48
$\Delta_{peak}$ , (mm)	48	48	36	36	36	24

Due to similar reasons, joints with higher prestress levels achieved greater stiffness as shown in Figure 10. Thus, they were able to attain their capacity without undergoing excessive deflection. On the contrary, the effect was completely reversed at the later stages of loading (after the 4<sup>th</sup> cycle). This was due to substantial loss of strength after peak response which resulted from prestress loss and was more pronounced for highly prestressed joints.

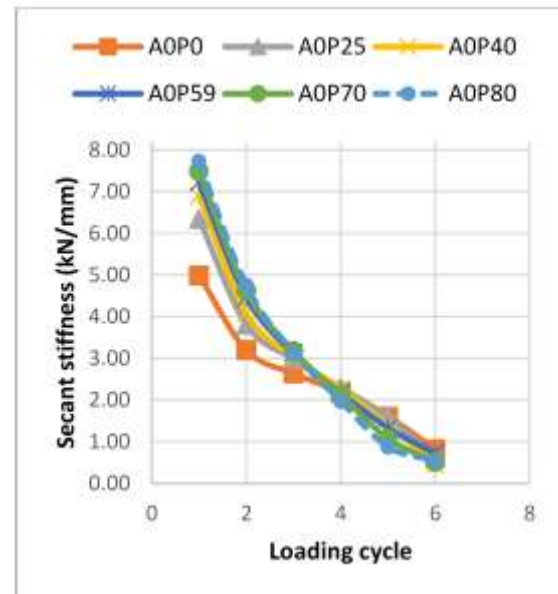


Figure 10 Secant stiffness of exterior joints, computed for each loading cycle

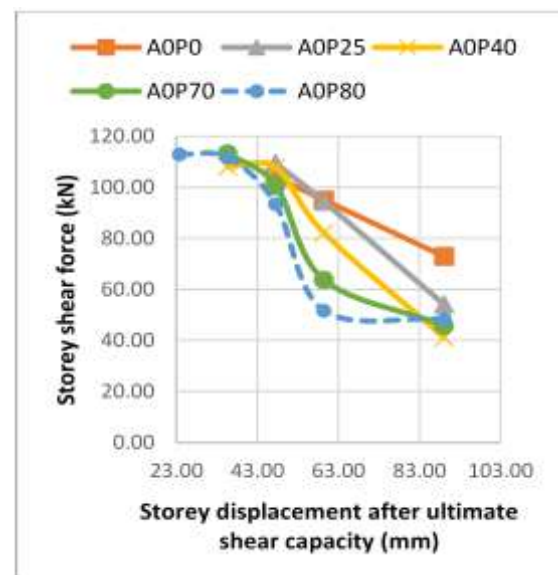


Figure 11 Strength degradation after peak response

The strength degradation after peak response and loss of prestress is shown in Figure 11 and Figure 12. These phenomena undermined the energy dissipation capacity of joints with higher prestress levels at the later cycles of the loading as shown in Figure 13.

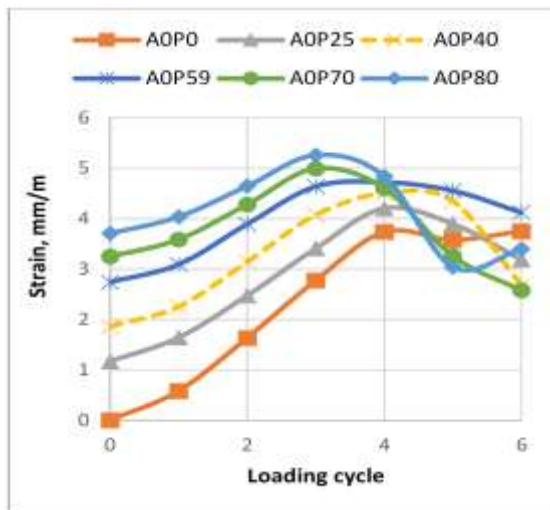


Figure 12 Strain distribution in the prestressing steel at the joint for each loading cycle

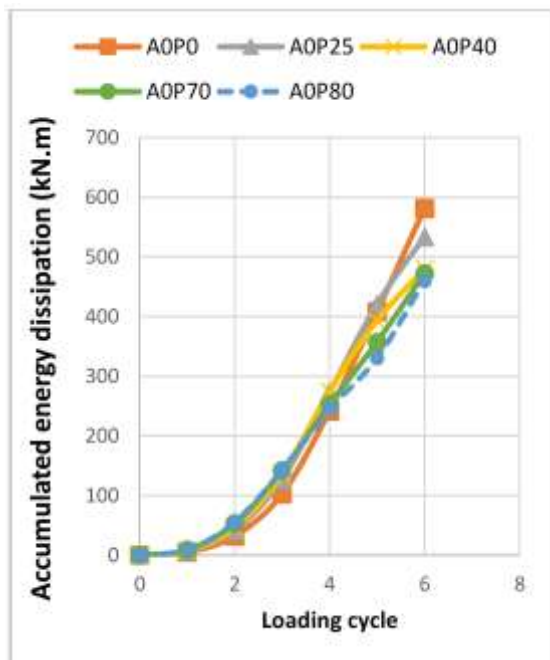


Figure 13 Effect of prestressing force on accumulated energy dissipation capacity

Effect of column axial load on the partially prestressed concrete Interior joints

The seismic performance of these specimens was evaluated based on premature crushing of joint concrete due to high compressive stress from the column axial load and prestressing steel. The

column axial load ratio effect was also evaluated on the Pinching behavior of the joints.

#### (a) Premature crushing of concrete joint

To study the premature crushing of concrete joint two parameters were taken into consideration. The first one is the ultimate storey shear capacity and the second one is an observation of internal vertical concrete stresses at the joint that is obtained at the last cycles of the loading in specimen IA65P80. Figure 14 clearly shows the effect of column axial load on the ultimate storey shear capacity of the specimens. Based on the analysis result, column axial load increment slightly enhanced the shear capacity of the joints. This is because the incoming compressive stress from the column axial load restrains the joint region against shear failure which considerably improves bond performance by preventing slippage of reinforcement bars.

In addition, an average concrete compressive stress observed in the beam-column joint at the final loading stage of specimen IA65P80 was 15MPa which is 56.3% of the compressive strength of the concrete (34.4MPa).

Thus, since no drop of ultimate strength was encountered with column axial load increment and the incoming stress is still much less than the compressive strength of concrete at the joint, crushing of concrete is not expected to occur. Therefore, premature crushing of concrete due to high compressive stress from the column axial load and prestressing steel did not take place.



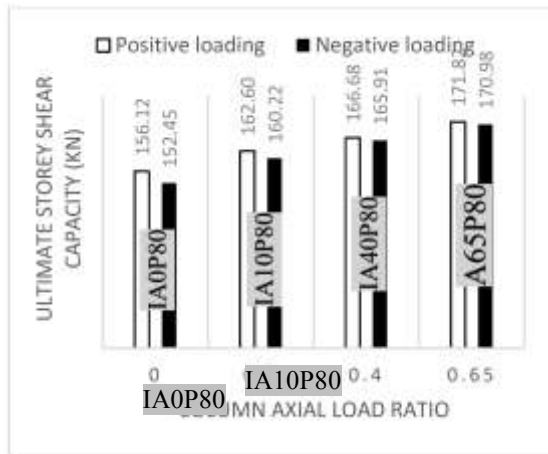


Figure 14 Effect of column axial load on the ultimate storey shear capacity

#### (b) Pinching

In reinforced concrete structures subjected to reverse cyclic loading, a large reduction in loading stiffness after unloading might occur. In the reloading branch of a hysteresis curve, after repeated cycles, cracks that occur previously in the tension side might still be open in addition to the one that is going to be created at the new face under tension. As a consequence, the concrete at the section will be fully cracked and the concrete will be ineffective in resisting the shear. This phenomenon causes narrowing of the hysteresis loops. Such effect is known as pinching. The column axial load tends to offset the pinching behavior of the interior joints. As shown in Figure 7, specimen IA0P80, with no column axial load, experienced severe pinching. But as the column axial load increases, a moderate pinching was observed in the specimens. In specimen IA40P80 and IA65P80, the high compressive stress from the column axial load delayed the formation of shear crack at the joints as can be seen from Figure 15. Therefore, since majority of these joint regions were not cracked immediate recovery of stiffness was possible.

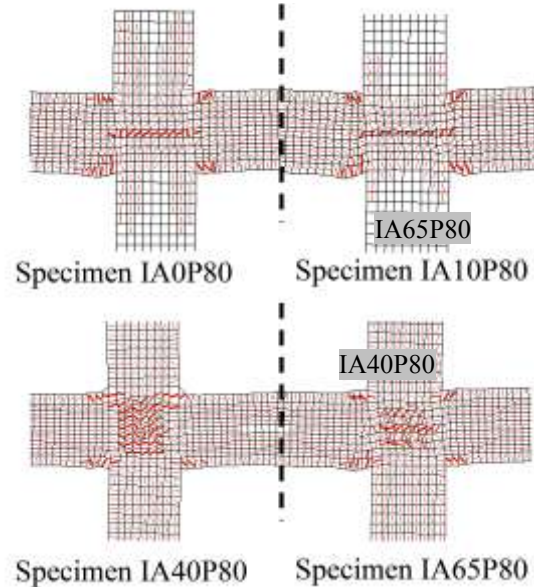


Figure 15 Failure crack pattern of the interior joint at the final stage of loading

## CONCLUSIONS

In this study analytical investigation on the effect of prestressing level and column axial load on the seismic performance of partially prestressed concrete exterior and interior beam-column joint was studied. According to the analytical result, variation of prestressing force did not encounter a significant effect on the ultimate shear capacity. Although, a significant increment in the ductility and stiffness of the joint was observed, this effect was completely reversed after the 4<sup>th</sup> loading cycle due to significant degradation of strength following the loss of prestress.

This phenomenon undermined the energy dissipation capacity of joints with higher prestress levels at the later cycles. On the other hand, in the interior joint specimens, column axial load ratio variation in a highly prestressed joint did not encounter



premature crushing of concrete at the joint at any level of the loading. Furthermore, a less pronounced pinching was observed with column axial load increment.

## REFERENCES

- [1] Nishiyama, M. "Seismic design of prestressed concrete buildings," Bulletin of New Zealand National Society for Earthquake Engineering, vol. 23, no. 4, pp. 288–304, 1990.
- [2] Kashiwazaki T. and Noguchi H., "Structural performances of prestressed concrete interior beam-column joints," Proc. 12th World Conference Earthquake Engineering (12WC EE), Auckland, New Zealand, January 30-February 4, 2000, pp. 1–8, 2000.
- [3] Paulay T. and Priestley M. J. N., "Seismic Design of Reinforced Concrete and Masonry Buildings," Second edition JOHN WILEY AND SONS, INC., 1992.
- [4] Berter V.V., and Shadh H., EI Asnam, Algeria Earthquake, October 10, 1980, Earthquake Engineering Research Institute, Oakland, Calif., January 1983, 190p.
- [5] Mitchell D., "Structural damage due to the 1985 mexico earthquake," Proceedings of the 5<sup>th</sup> Canadian Conference on Earthquake Engineering, Ottawa, 1987, A. A. Balkema, Rotterdam, pp. 87-111.
- [6] Caltrans Seismic Design References, California Department for Transportation, Sacramento, Calif., June 1990.
- [7] Loma Prieta Earthquake October 17, 19 Preliminary Reconnaissance Report, Earthquake Engineering Research Institute, Oakland, Calif., 50 p.
- [8] Park R. and Paulay T., "Reinforced concrete structures," First edition JHON WILEY & SONS, INC., 1975.
- 9] Nishiyama M. and Wei Y., "Effect of post-tensioning steel anchorage location on seismic performance of exterior beam-to-column joints for precast, prestressed concrete members," PCI Journal, vol. 52, pp. 18–30, 2007.
- 10] Wong, P. S. Vecchio F. J., and Trommels, H. "VecTor2 and formworks user's manual," Second Edition 2013.
- [11] CBE-F IP Model Code 90, Institute of Civil Engineers (ICE) Publication., pp. 82-116



# MECHANICAL AND DURABILITY PROPERTIES OF CONCRETE WITH GRANITE POWDER AS PARTIAL REPLACEMENT OF CEMENT

Asnake Kefelegn<sup>1</sup>, Binaya Patnaik<sup>2</sup> and Issayas Kebede<sup>3\*</sup>

<sup>1</sup>Department of Civil Engineering, Hawassa University, Hawassa, Ethiopia

<sup>2</sup>Department of Civil Engineering, Gambella University, Gambella, Ethiopia

<sup>3</sup>Construction Supervision Team in SNNPR's Construction Authority, Hawassa, Ethiopia

\*Corresponding author. E-mail address: [issayasworku@gmail.com](mailto:issayasworku@gmail.com)

## ABSTRACT

*This experimental investigation was performed to evaluate the strength and durability properties of concrete, in which cement was partially replaced with granite powder at 5%, 10%, 15%, and 20% by volume. A total of five concrete mix proportions were developed for control and concrete with granite powder. The fresh property test, strength, and durability tests were done for normal strength concrete (NSC) as well as high strength concrete (HSC). The fresh property of concrete was evaluated using slump test, and the strength of concrete was evaluated using compressive strength and splitting tensile strength test. Durability properties were evaluated using water absorption test, sorptivity test, and chloride and sulphate attack test for all concrete mixes. From the investigation, the compressive strength was enhanced up to 10% replacement in NSC and at 5% replacement in HSC. The tensile strength was improved up to 10% replacement in the both types of the concrete. Denser concrete against water absorption was made at 5% & 10% in NSC and 5% in HSC. The initial rate of water absorption (sorptivity) was increased at 5% in NSC and 10% in HSC concrete. The resistance to chloride and sulphate attack was enhanced at all replacement percentages for NSC as well as HSC.*

**Keywords:** Cement replacement, durability properties, Granite powder, Mechanical properties

## INTRODUCTION

Portland cement is hydraulic cement mainly composed of hydraulic calcium silicates. Hydraulic cement is set and hardened by reacting chemically with water. During this reaction, called hydration, cement combines with water to form a stone-like mass, called paste. When the paste (cement and water) is added to aggregates, it acts as an adhesive and binds the aggregates together to form concrete, the world's most versatile and most widely used construction material [1].

The following are the four primary compounds in Portland cement, their approximate chemical formulas, and abbreviations: Tricalcium silicate  $3\text{CaO}\cdot\text{SiO}_2 = \text{C}_3\text{S}$ , Dicalcium silicate  $2\text{CaO}\cdot\text{SiO}_2 = \text{C}_2\text{S}$ , Tricalcium aluminate  $3\text{CaO}\cdot\text{Al}_2\text{O}_3 = \text{C}_3\text{A}$ , Tetra calcium aluminoferrite  $4\text{CaO}\cdot\text{Al}_2\text{O}_3\cdot\text{Fe}_2\text{O}_3 = \text{C}_4\text{AF}$  [1, 2]. In addition to the four major compounds, there are many minor compounds formed in the cement. The influence of these minor compounds on the properties of cement or hydrated compounds is not significant [2].

The following are the approximate equations of the reactions of  $\text{C}_3\text{S}$  and  $\text{C}_2\text{S}$  with water.

$2(3\text{CaO}\cdot\text{SiO}_2) + 6\text{H}_2\text{O} \rightarrow 3\text{CaO}\cdot 2\text{SiO}_2\cdot 3\text{H}_2\text{O} + 3\text{Ca}(\text{OH})_2$ . Similarly,  $2(2\text{CaO}\cdot\text{SiO}_2) + 4\text{H}_2\text{O} \rightarrow 3\text{CaO}\cdot 2\text{SiO}_2\cdot 3\text{H}_2\text{O} + \text{Ca}(\text{OH})_2$ .  $\text{Ca}(\text{OH})_2$  is not a desirable product in the concrete mass.

It is soluble in water and gets leached out making the concrete porous, particularly in hydraulic structures. And, the lack of durability of concrete is on account of the presence of calcium hydroxide. The calcium hydroxide also reacts with sulphates present in soils or water to form calcium sulphate which further reacts with  $C_3A$  and causes deterioration of concrete. This is known as sulphate attack. To reduce the quantity of  $Ca(OH)_2$  in concrete and to overcome its bad effects by converting it to cementitious products is the advancement in concrete technology. The use of blending materials such as fly ash, silica fume, and other pozzolanic materials are the steps to overcome the bad effect of  $Ca(OH)_2$  in concrete [2].

For reinforced concrete structures, especially for water retaining structures, the limiting of crack width as a result of shrinkage is important. Thermal shrinkage can be reduced by restricting the temperature rise during hydration, which can be achieved by the mix proportions with low cement content or suitable cement replacement e.g. fly ash (pulverized fuel ash) or ground granulated blast furnace slag. Cement containing ground granulated blast furnace slag or fly ash will not only help to reduce temperature rises due to hydration but will also increase durability [3].

Granite powder is a by-product produced in granite factories while cutting huge granite rock to the desired shapes. This granite powder has a chemical composition like the raw materials used for manufacturing cement [4]. Based on ASTM C618, if the sum of the percentage composition of silica, alumina, and ferric oxide is over 70%, the material can be introduced to concrete as a pozzolanic material [5]. The effect of replacing granite fines as the sand on vibrated structural concrete has been studied by different researchers [6, 7, 8, 9, 10].

Divakar Y., et al. used a concrete which was prepared with granite fines as a replacement of fine aggregate in 5 different proportions namely 5%, 15%, 25%, 35%, and 50%, and various tests such as compressive strength, split tensile strength and flexural strength were conducted and these test values were compared with the conventional concrete without granite fines. In this investigation, the compressive strength was increased by 22% with 35% replacement of fine aggregate with granite fines, and the compressive strength was still higher than the control's samples strength for up to 50% replacement. At 50% replacement of granite fines, the compressive strength was 38.5 MPa whereas the control was 37 MPa. The splitting tensile strength was not significantly affected up to 50% replacement. The flexural strength of 10cm x 10cm x 50cm prism without reinforcement increased at 5% replacement by 5.41%, but its value decreased with the replacement beyond 5% even if the reduction was insignificant. The flexural strength of 15 cm x 15 cm x 70 cm beam with reinforcement was checked at 25% and 50% replacements, and the result showed that at 25% replacement a 2% increment was observed and at 50% replacement the strength was increased by 32.7% [6].

Raja G. & Ramalingam K. investigated the mechanical properties of normal-strength concrete by replacing sand with granite fines. The percentage replacement of granite fines used were 10, 20, 30, 40, 50 & 100 for concrete cube strength of 20 MPa mix proportions. Specimens were tested after 28 days of curing for compression strength, flexural strength, and tensile splitting test. From the study, the specimens with 40% replacement of granite fines achieved higher strength compared to the control specimen [7]. Allam M., et al. investigated the effect of replacing the sand with granite waste in the concrete mix at the values of 10%, 17.5%, and 25%.

In this study, splitting tensile strength after 28 days of curing was increased by 12%, 15%, and 21% respectively as compared to the control mix. By replacing the sand with 10% granite granules by weight, the value of the flexural strength was increased by 34% and at 17.5% replacement, the value dropped back to the same as that of the control. At the highest — 25% percentage of replacement, the value of flexural strength was 15% lower than the control mix. By replacing the 10% sand with granite powder, the value of bond strength increased by 1%—further increase decreases the bond strength [8].

Shehdeh G., et al. investigated the effect of replacing granite powder and iron powder as sand at 5%, 10%, 15%, and 20% by weight. From the investigation, it was observed that substitution of the 10% of sand by weight with granite powder in concrete was the most effective in increasing the compressive and flexural strength compared to other replacement percentages. The test result showed that for a 10% ratio of granite powder in concrete, the increase in the compressive strength was about 30% compared to the control samples. Similar results were observed for the flexural strength. It was also observed that substitution of up to 20% of sand by weight with iron powder in concrete resulted in an increase in compressive and flexural strength of the concrete [9].

Shinde S., et al. investigated the effect of sand replacement with granite powder at 10%, 20%, 30%, 40%, 50% & 100%. In this study, the effect on the compressive and tensile strength was examined. The result from the study showed that the maximum compressive and tensile strength was attained at 20% replacement of granite powder. Up to 50% replacement, the compressive strength was higher than the compressive strength of the control samples. And up to 40%

replacement, the tensile strength was also higher than the controls [10]. A review on partial replacement of cement material in Ethiopia has been carried out in Makebo G., et al. [11]. In this review work, it was stated that the waste materials like coffee husk ash, banana leaf ash, bagasse ash, bone powder, corncobs ash, municipal waste, coal mine, lime sludge, groundnut shell ash, quarry dust, and iron tailing have pozzolanic properties and can partially replace cement in the range of 10% – 15% in medium strength concrete production. The optimum percentage replacement of the materials was 10%. And, if the percentage replacement of the materials increases, the compressive strength starts decreasing.

The effect of replacing cement with granite powder on vibrated structural concrete was investigated by different researchers [8, 12, 13, 14].

The splitting tensile test on the concrete cylinders with different proportions of granite waste as partial replacement of cement was studied in Allam M., et al. (2016) [8]. In this study, it was shown that at 5% of granite fines waste as a partial replacement of cement, the strength was 20% higher than the control mix, but at 10% replacement, the strength dropped to the value equal to the control. In contrast, the flexural strength of the mixes containing 5%, 10%, and 15% of fine granite waste as a partial replacement of cement was 19%, 30%, and 37% lower than the control mix respectively. The bond strength of mix containing 5% of fines as replacement of cement was slightly higher by 1% [8].

Koneti V., et al. used granite slurry and sawdust to partially substitute cement and sand respectively. Sawdust replaced the fine aggregate at 3%, 5%, and 7% whereas granite slurry replaced the cement by 10%, 20%, and 30%. At 10% of cement



replacement with the granite slurry, the corresponding saw dust replacement was 3%. Similarly, at 20% replacement of cement with granite slurry, the corresponding saw dust replacement was 5% and for 30% cement replacement by granite slurry, sawdust replaced sand at 7%. The result from the investigation showed that the compressive strength on the seventh day was almost two times greater than the control mix in all replacement of granite and sawdust which indicates improved early strength gain. The maximum compressive strength was attained at 10% replacement of granite slurry and at 3% replacement of sawdust. Similarly, at 10% replacement of cement with granite slurry and 3% replacement of sand with sawdust, the maximum tensile strength value was attained [12].

Chiranjeevi R., et al studied the strength properties of concrete by using granite powder as an admixture. Concrete with cubic compressive strength of 25 MPa was prepared with the granite fines as a replacement of cement in concrete at different proportions namely 2.5%, 5%, 7.5%, 10%. From the investigation, at the optimum 7.5% replacement of cement by granite waste, the maximum compressive strength with a percentage increment of 42.47% was attained. The splitting tensile strength and the flexural strength were also maximum at 7.5% replacement of cement with granite powder [13].

The investigation of the fresh and hardened properties of ready mix concrete by partial replacement of cement with granite powder, and using manufactured sand and super plasticizer was carried out by Srinivasa C., et al. (2009)[14]. In this investigation, the partial replacement of ordinary portland cement with granite powder by 10%, 15%, 20%, 25%, 30%, 35%, and 40% was carried out. From the investigation, it was observed that the workability and compaction factor were acceptable for all mix batches which

satisfy the requirements of ready-mix pumpable concrete. The compressive strength at 28 days with 20% replacement was the maximum one from which the optimum percentage was established for the target mean strength value.

The durability of concrete made with granite powder replacing cement was studied in Allam M., et al in this study, the Scanning Electron Microscope (SEM) of concrete images for 5% granite waste powder as a partial replacement of cement indicates a denser concrete mix with the lowest number of pores. Additionally, Bakhoum E., et al. found the durability improvement of a mortar. In this study, the SEM images of the mix containing replacement of 5% cement and 10% sand with nano-granite waste showed maximum density and minimum micro-cracks and number of pores [16].

As it is observed from different investigations reviewed above, the optimum cement replacement percentage with granite powder for normal strength concrete varies from 5% to 20%. Moreover, in the review part of this article, almost all of the studies were on strength properties and durability cases were not investigated in detail. Furthermore, lots of effort has been done on investigating the strength properties of concrete using granite waste as a partial replacement of fine aggregate. And a few researches were performed on the strength properties of medium-strength concrete by replacing cement partially with granite slurry.

In this study, the granite powder used was finer than the powder used by other researchers, and the cement replacement by volume was also adopted. The investigation also included the granite powder's replacement effect on high strength concrete and durability in addition to the investigation of its effect on medium strength concrete.

## MATERIALS

In this experimental study, the medium strength concrete, C20/25 (NSC,) and high strength concrete, C55/67 (HSC) were used [17]. The concrete test specimens were cast by replacing cement with granite powder at 5%, 10%, 15%, and 20% by volume and cured for strength and durability property investigations. The cement used was Dangote Ordinary Portland Cement (OPC) with a 42.5R grade. The fine and coarse aggregates used were locally available materials that were collected from Dimtu and Monopole around Hawassa city respectively. The physical properties of used sand and coarse aggregates which were determined as per the manual [18] are put in Table 1 and 2 respectively. The maximum coarse aggregate size used for medium and high-strength concrete was 25mm and 19mm respectively. The bulk unit weights were also  $1372 \text{ kg/m}^3$  and  $1360 \text{ kg/m}^3$  respectively.

*Table 1. Physical properties of aggregates*

Fineness modules, FM	2.81
Silt Content, %	3.57
Specific Gravity (OD)	2.33
Absorption, %	2.04
Moisture content, %	2.04

*Table 2. Physical properties of coarse aggregates*

Specific gravity (OD)	2.55
Absorption, %	1.42
Moisture content, %	0.50

The granite powder used (shown in Figure 1), which was collected from COA General Trading PLC's workshop in Addis Ababa around Balderas signal, was finer than  $45\mu\text{m}$  (No 325) sieve and its chemical composition, which was tested in the Geological Survey of Ethiopia laboratory, is shown in Table 3.

The properties of the super plasticizer used in this investigation, which was taken from SAS Construction Chemicals Ltd's profile, are shown in Table 4.



Figure 1. Granite powde

Table 3 Chemical composition of granite powder

Chemical oxides composition	Percentage by weight
Silica (SiO <sub>2</sub> )	69.12
Alumina (Al <sub>2</sub> O <sub>3</sub> )	17.77
Iron (Fe <sub>2</sub> O <sub>3</sub> )	2.17
Calcium oxide (CaO)	1.54
Magnesia (MgO)	0.46
Soda (Na <sub>2</sub> O)	2.22
Potassium Oxide (K <sub>2</sub> O)	3.86
Manganosite (MnO)	0.04
Potassium oxide (P <sub>2</sub> O <sub>5</sub> )	0.05
Titanium dioxide (TiO <sub>2</sub> )	0.14
Water (H <sub>2</sub> O)	0.1

Table 4. Properties of super plasticizer

Properties	Observations
Colour	Dark brown liquid
Specific gravity	1.22 ± 0.03 at 25°C
Chemical base	Naphthalene sulphonate
Air entrainment	1-2 % depending on dosage
Chloride content	Nil

## 1. EXPERIMENTAL METHODS

### 1.1. Mix design, mixing, and curing procedures

The mix proportions for medium and high strength concrete which was designed as per ACI 211.1-91 [19] and ACI 211.4R -93 [20] respectively are summarized in Table 5. Hand mixing and tamping of the fresh concrete in the standard mold were carried out as per ASTM C 192 M-02 standards [21]. All specimens were moist cured at room temperature from the time of molding till the moment of the test as per ASTM C 192M-02 standard [21].

Table 5. Mix proportions of medium and high strength concrete

Materials	C20/25	C55/67
	kg/m <sup>3</sup>	kg/m <sup>3</sup>
Water	201.52	217.88
Cement	327.81	733.18
Coarse aggregates	937.34	984.17
Fine aggregates	774.22	382.34
Super plasticizer	0	11

### 1.2. Tests carried out

The slump test for workability was carried out as per ASTM C 143/C 143M - 00 standard [22] for each case specimen both for medium and high strength concrete. The compressive strength of cube 15cm-size concrete specimens was tested as per ASTM C 39/C 39M standard [23] for each granite powder replacement case and both for medium and high-strength concrete. Three test specimens were tested for selected curing ages, 7<sup>th</sup> and 28<sup>th</sup> days, of concrete. Each compressive strength specimen was subjected to a 0.4 MPa/sec loading rate.

The flexural strength of 15x15cm cross-section size with 50cm span length plain concrete specimens was tested as per ASTM C 293 - 02 standard [24] for each granite powder replacement case and both for medium and high strength concrete. In this test, two test specimens were tested for the only 28<sup>th</sup> day of concrete. The flexural strength specimens were subjected to a 0.02 MPa/sec loading rate.

The water absorption by immersion test was done based on ASTM C 642 - 97 standard [25]. The water absorption of three cubes of 15 cm size was tested for each granite powder replacement case and both for medium and high 28<sup>th</sup>-day strength of concrete.

Sorptivity measures the rate of water absorption of hydraulic cement concrete by measuring the increase in the mass of a specimen resulting from absorption of water as a function of time when only one surface of the specimen is exposed to water. The initial rate of water absorption (sorptivity) is the absorption from one minute to six hours. In this study, the sorptivity test was carried out as per ASTM C 1585 - 04 standard [26]. Two-disc slices of the concrete cylinder for each granite powder replacement case were



used in this study. The slices used for this test were the middle two slices after rejecting the top and bottom disc slices. Slice specimens of 5 cm depth were

prepared by cutting a cylinder concrete specimen with the size of 5 cm diameter and with a depth of 20 cm into four equal parts as shown in Figure 2.



Figure 2. Sorptivity test specimens after cutting cylinder.

The test method for the chemical resistance of concrete is specified in the ASTM C 1012 [27]. However, this method is for mortar, and the behavior of mortar and concrete under chemical attack might not be the same. One of the ways of knowing the deterioration mechanism of concrete under the exposed chemical solution is the mass loss method as shown by equation 1 [28].

In order to get the accelerated degradation process and to shorten the test duration, changing the concentration of the sulfate solution in a way that simulates the highest sulfate concentrations can be done [29]. The lower limit of the exposure proposed by the ASTM C1012 [27] test method is the use of 50,000 mg/L  $\text{Na}_2\text{SO}_4$  concentration in water solution.

In this sulfate and chloride chemical attack experimental study, JSTM C7401 [30] testing method is used. This test method assesses the chemical resistance of concrete by immersing test specimens into acid or alkaline solutions for a prescribed period and by comparing changes in the measurements against control specimens. The sulfate resistance of concrete can be quantified by measuring changes in weight

occurring in specimens stored in chemical solutions [31].

In this experimental investigation, cubes of concrete of 15cm, which were cured for 28 days for both normal and high strength concrete, were used. After the final day of curing, the specimens were removed from the water, and the excess film of water on the surface was cleaned using a standard preliminary surface cleaning process and weighed as initial mass. Then the identified specimens were immersed in the 5% sulfate and chloride chemical solutions for another 28 days. After the prescribed duration, the specimens were removed from the solution and their final weight was recorded. Then, sulfate and chloride resistance of the specimens in terms of weight loss was determined using equation

$$\text{Mass change (\%)} = \frac{w_o - w_f}{w_o} \times 100\% \quad (1)$$

Where  $W_f$  is the mass of concrete immersed in a test solution on the 28<sup>th</sup> day (g), and  $W_o$  is the mass of concrete before immersion in a test solution (g).

## 2. RESULTS AND DISCUSSION

In this section, the results from the experiment and discussion are presented. The test result for the slump value is shown in Figure 3.

As observed from Figure 3, the slump value for both types of concrete decreased, which may happen from the higher surface area of granite powder which can increase the surface of hydration leading to higher water absorption. The addition of powder was also observed to result in loss of slump as reported in [32] for medium-strength concrete.

The test result for average compressive strengths is shown in Figure 4. From the test results, for NSC concrete, the 7<sup>th</sup>-day

strength test result for 5% and 10% replacement of cement with granite powder increased by 13% and 9% respectively. Whereas for HSC concrete, the 7<sup>th</sup>-day strength test result increased at 5% replacement by 5.86%. However, it decreased for 10%, 15%, & 20% replacement compared the control strength.

For NSC concrete, the 28<sup>th</sup>-day compressive strength test result showed that the average compressive strength at 5% and 10% replacement was higher by 3.36% and 1% respectively. However, relative to the 7<sup>th</sup>-day strength result, the 28<sup>th</sup>-day strength increment is lower. This indicates that partial replacement of cement with granite powder improved an early strength gain.

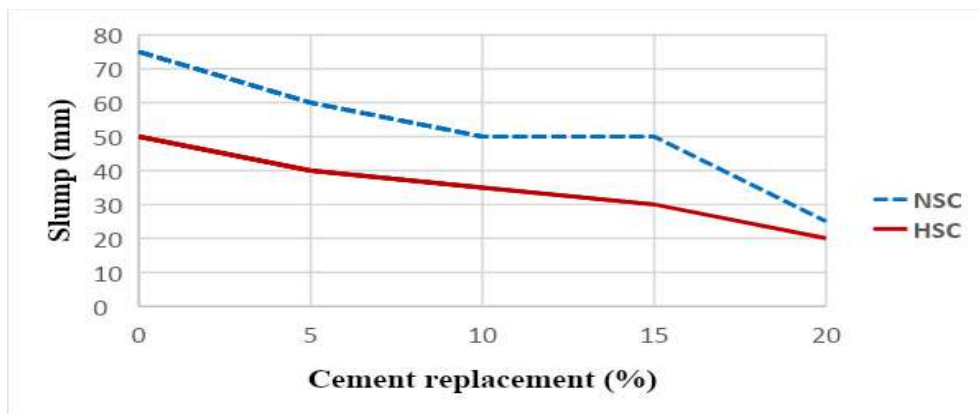


Figure 3. Slump value for each replacement

The strength result for HSC concrete showed that concrete cubes containing 5% granite powder are higher than the control strength both on the 7<sup>th</sup> and 28<sup>th</sup> days. Its strength is increased by 5.86% and 6.78% on the 7<sup>th</sup> and 28<sup>th</sup> day respectively. For the other replacements, the strength is decreased. The reason for the enhancement of the strength may be fine powders chemically react with calcium hydroxide at ordinary temperatures to form compounds having cementitious properties. When using these materials in concrete, the concrete will make efficient use of the hydration products of Portland cement and

consume calcium hydroxide to produce additional cementing compounds.

The test result for average flexural strengths is shown in Figure 5. From the result for NSC concrete, it is observed that concrete beams containing 5% and 10% granite powder attained greater flexural strength compared to the control by 6.34% and 7.94% respectively. However, it decreased by 1% and 1.12% at 15% and 20% replacement respectively. For HSC concrete, the flexural strength is also enhanced up to 10% replacement of cement with granite powder. At 5%

replacement, the flexural strength increased by 6.24%, and at 10% replacement, the strength increased by 4.90% compared to the control beams. But, for 15 & 20% replacement, the strength decreased by 1.32% & 10.52% respectively.

The average water absorption by weight is shown in Figure 6. From this figure, for the NSC concrete, the percentage of water

absorption by weight decreased for 5% and 10% replacement. For HSC concrete, the water absorption performance was improved for concrete containing 5% granite powder. This is probably due to the filling effect of granite micro-sized particles which reduced the volume and conductivity of capillary pores which in turn creates fewer voids to permit water to go through.

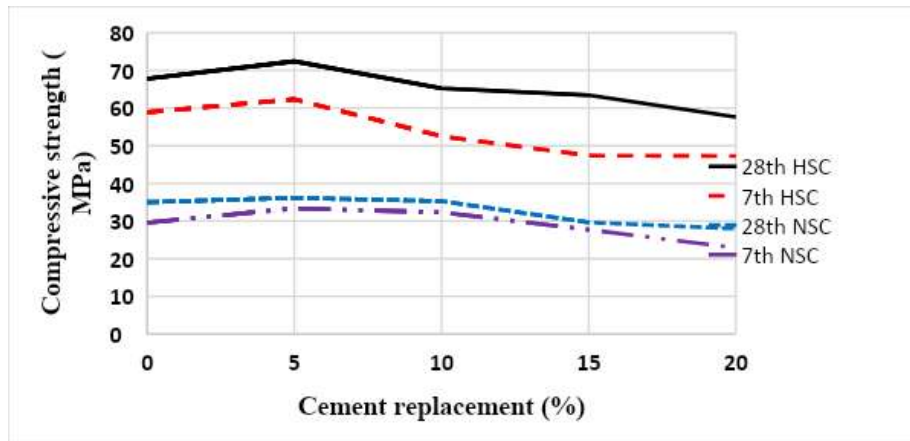


Figure 4. Compressive strength of concrete for each replacement

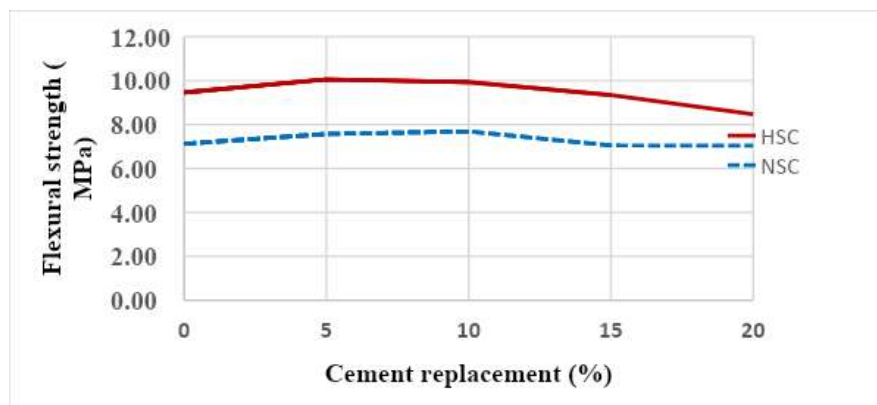


Figure 5. Flexural strength of concrete for each replacement

The absorption ( $I$ ) is the change in mass divided by the product of the cross-sectional area of the test specimen and the density of water. The initial rate of water absorption (sorptivity) is defined as the slope of the line that best fits to  $I$  plotted against the square root of time ( $s^{1/2}$ ) between one minute and six hours. Moreover, according to the ASTM C 1585 – 04[26] standard, the result is valid only

for a correlation coefficient greater than 0.98. Otherwise, the result is no longer representative, and hence, the rate of water absorption cannot be determined. The result for the initial rate of water absorption test is shown in Figure 7. From this figure, for NSC concrete, it is observed that the initial rate of water absorption enhanced at 5% replacement. For HSC concrete, the enhancement was

observed at 10% replacement. And, the observed water absorption rate behavior follows a contrasting pattern i.e. when the

NSC's absorption rate decreases, HSC's absorption rate increases.

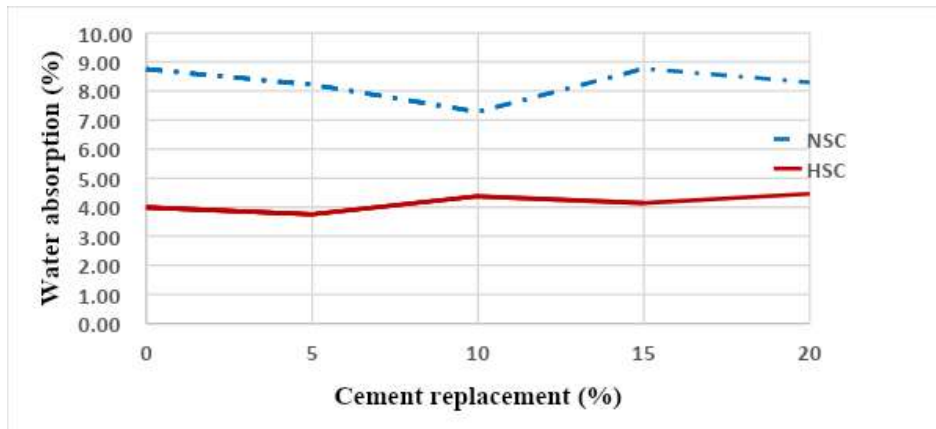


Figure 6. Water absorption of concrete for each replacement

The result for the sulphate and chloride attack test is shown in Figure 8. An “S” and “C” letters are added to the acronym NSC and HSC to show the test result for sulfate and chloride chemical solutions respectively. From Figure 8, it is observed that for all replacements of granite powder, the resistance of concrete was improved under all chemical solutions. The optimum percentage of replacement against sulfate and chloride attack resistance was attained at 5% and 10% replacement for NSC concrete and 5% replacement for chloride attack and 5% and 10% replacement for sulfate attack for HSC concrete.

The reason behind the improvement against sulphate attack might be sulphate salt attacks either  $C_3A$ , calcium hydroxide (CH), or mono sulfoaluminate (AFm). Then it forms ettringite which is expansive and causes a crack. Once the salt has consumed all the CH, then it starts to decalcify calcium silicate hydrate (CSH) which is the backbone of concrete strength.

Chloride dissolved in waters tends to increase the rate of leaching of port landite ( $Ca(OH)_2$ ), thus it increases the porosity of concrete and loses weight. [31]. The rate of ingress of chlorides and penetrability of concrete depends on the pore structure of the concrete, which is affected by materials used in the concrete. This will be influenced by the water to cement (w/c) ratio of the pozzolanic materials which serve to subdivide the pore structure [33].

As a result, the improvement against chloride attack is observed for all replacement of granite powder in both NSC and HSC concrete (Figure 8). Moreover, the weight loss observed in HSC concrete was lower than NSC concrete, which may happen due to the presence of denser structure or lower voids in HSC concrete than NSC as a result of the lower water to cement (w/c) ratio used in HSC.

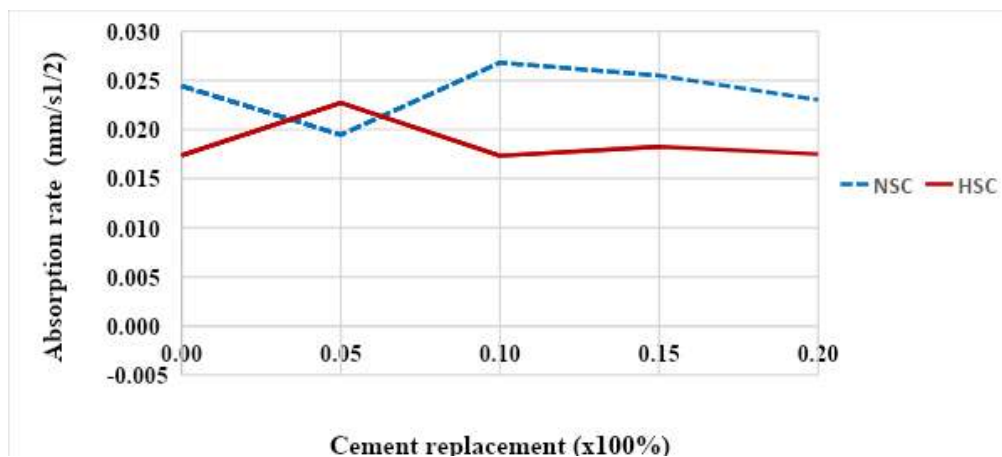


Figure 7. Initial rate of water absorption

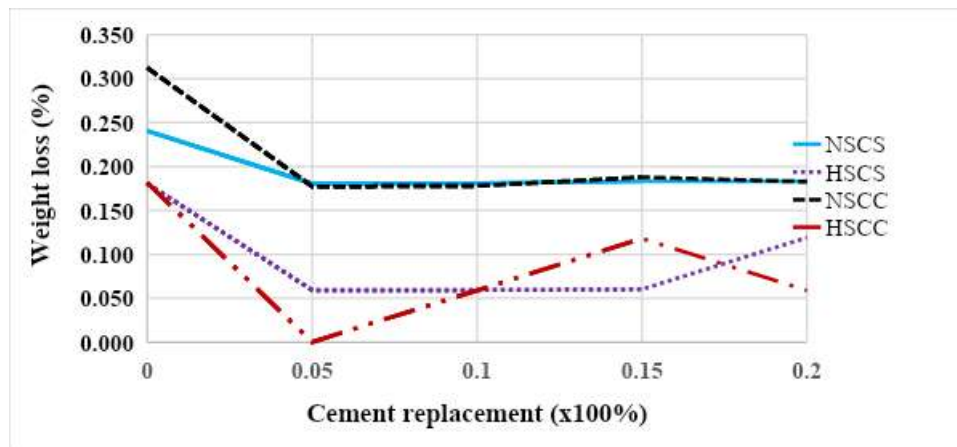


Figure 8. Sulphate and chloride attack (weight loss).

## CONCLUSIONS

In this study, the effect of replacing cement partially with granite powder at 5%, 10%, 15%, and 20% were investigated experimentally. The investigation included workability of fresh concrete using slump value, the strength of concrete using compressive and flexural tensile strength, and durability using water absorption test, sorptivity test, and chloride and sulphate attack test for both NSC and HSC concrete specimens and for each granite powder replacement. From the results of the investigation, the following conclusions are drawn.

1. The workability of both NSC and HSC concrete decreased in a linear manner as the percentage replacement increased. This may happen as a result of the higher surface area of granite powder which increased the surface of hydration leading to higher water absorption.
2. Granite powder enhanced the NSC concrete compressive strength of the 7<sup>th</sup> day by 13 % and 9 % and the 28<sup>th</sup> day by 3.36 % and 1 % at 5 % and 10 % replacements respectively. Moreover, the 5% granite powder replacement enhanced the compressive strength of HSC concrete by 5.86 % and 6.78 % on the 7<sup>th</sup> and 28<sup>th</sup> day respectively. For the other replacements, the strength is decreased.

The reason for the enhancement may be granite powder chemically react with calcium hydroxide at ordinary temperatures to form compounds having cementitious properties which give additional strength to the concrete.

3. Similarly, the flexural tensile strength was enhanced at 5 % and 10 % replacements for both NSC and HSC concrete. For NSC concrete, it is observed that concrete beams containing 5% and 10% granite powder attained greater flexural strength compared to the control by 6.34% and 7.94% respectively. Similarly, for HSC concrete, the flexural strength increased by 6.24% at 5% replacement and by 4.90% at 10% replacement compared to the control beams. For the other replacements, the strength is decreased.
4. A denser and least permeable concrete with the least water absorption is made at 5% and 10% replacement for NSC and 5% replacement for HSC. This happened might be due to the filling effect of granite powder in which fewer conductivity voids are made at these percentages of the replacements.
5. The least initial rate of absorption is observed at the 5% replacement for NSC concrete and 10% replacement for HSC concrete. Hence, it can be concluded that granite powder had a significant effect in reducing the water absorption by capillary suction.
6. The weight loss resistance of concrete to chloride and sulphate attack was enhanced in all replacements relative to the control specimen for both NSC and HSC concrete. The optimum percentage of replacement against sulfate and chloride attack was attained at the 5% and 10% replacement in NSC concrete and 5% replacement for chloride attack and 5% and 10% replacement for sulfate attack in HSC concrete.

## ACKNOWLEDGMENTS

This study was done at Hawassa University Institute of Technology collaborating with Ethiopia Roads Authority. Hence, the authors need to acknowledge these institutions for their backing.

## REFERENCES

- [1] Kosmatka, H.S., et al., "Design and Control of Concrete Mixtures", 14<sup>th</sup> Ed., Portland Cement Association, 2002.
- [2] Shetty, M. S. and Jain, A. K., "Concrete Technology (Theory and Practice)", 8<sup>th</sup> Ed., Chand Publishing, 2019.
- [3] Mosley, B., et al., "Reinforced Concrete Design to Eurocode 2", 7<sup>th</sup> Ed., Palgrave Macmillan, 2020.
- [4] Patel, A. J., et al., "Review on Partial Replacement of Cement in Concrete", UKIERI Concrete Congress–Concrete Research Driving Profit and Sustainability, vol. 1, no. 7, 2015, pp. 831–837.
- [5] ASTM C 618 - 13, "Standard Specification for Coal Fly Ash and Raw or Calcined Natural Pozzolan for Use in Concrete", ASTM International, 2013.
- [6] Divakar, Y., et al., "Experimental Investigation on Behaviour of Concrete with the Use of Granite Fines", International Journal of Advanced Engineering Research and Studies, vol. 1, no. 4, 2012, pp. 84–87.
- [7] Raja, G. and Ramalingam, K., "Experimental Study on Partial Replacement of Fine Aggregate by Granite Powder in Concrete", International Journal for Innovative



- Research in Science & Technology, vol. 2, no. 12, 2016.
- [8] Allam, M., et al., “*Influence of Using Granite Waste on the Mechanical Properties of Green Concrete*”, ARPN Journal of Engineering and Applied Sciences, vol. 11, no. 5, 2016, pp. 2805–11.
- [9] Shedder, G., et al., “*Experimental Study of Concrete Made with Granite and Iron Powders as Partial Replacement of Sand*”, Sustainable Materials and Technologies, vol. 9, 2016, pp. 1 - 9.
- [10] Shined, S., et al., “*To Study the Effect of Granite Powder on Concrete*” International Journal for Research in Applied Science & Engineering Technology, vol. 6, no. 3, 2018.
- [11] Makebo, G., “*Partial Replacement of Cement Material in Ethiopia: A Review*”, International Research Journal of Engineering and Technology (IRJET), vol. 6, no. 1, 2019, pp. 1002 - 05.
- [12] Vamsi, K., et al., “*Partial Replacement of Cement in Concrete with Granite Powder and Fine Aggregate with Saw Dust*” International Research Journal of Engineering and Technology, 2019.
- [13] Chiranjeevi, K., et al., “*Experimental Study on Concrete with Waste Granite Powder as an Admixture*”, Int. J. Eng. Res. Appl, vol. 5, 2015, pp. 87–93.
- [14] Srinivasa, H., and Venkatesh, “*Optimization of Granite Powder used as Partial Replacement to Cement in the Design of Ready Mix Concrete of M20 Grade Using IS10262:2009*”, International Journal of Engineering Research & Technology, vol. 4, no. 1, 2015.
- [15] Allam, M., et al., “*Durability of Green Concrete Containing Granite Waste Powder*”, Int. J. Eng. Technol, vol. 8, no. 5, 2016, pp. 2383–2391.
- [16] Bakhoun, E. S., et al., “*The Role of Nano-Technology in Sustainable Construction: A Case Study of Using Nano Granite Waste Particles in Cement Mortar*”, Engineering Journal, vol. 21, no. 4, 2017, pp. 217–227.
- [17] Eurocode 2, “*Design of Concrete Structures - Part 1-1: General Rules and Rules for Buildings*”, European Committee for Standardization, 2004.
- [18] Dinku, A., “*Construction Materials Laboratory Manual*”, Addis Ababa University Printing Press, 2002.
- [19] ACI 211.1-91, “*Standard Practice for Selecting Proportions for Normal, Heavyweight, and Mass Concrete*”, American Concrete Institute, 2002.
- [20] ACI 211.4R-08, “*Guide for Selecting Proportions for High-Strength Concrete Using Portland Cement and Other Cementitious Materials*”, American Concrete Institute, 2008.
- [21] ASTM C192/C192M, “*Standard Practice for Making and Curing Concrete Test Specimens in the Laboratory*”, ASTM International, 2014.
- [22] ASTM C 143/C 143M, “*Standard Test Method for Slump of Hydraulic-Cement Concrete*”, ASTM international, 2000.
- [23] ASTM C 39/C 39M - 01, “*Standard Test Method for Compressive Strength of Cylindrical Concrete Specimens*”, ASTM international, 2001.

- [24] ASTM C 293 - 02, "*Standard Test Method for Flexural Strength of Concrete*", ASTM international, 2002.
- [25] ASTM C 642 – 97, "*Test Method for Density, Absorption, and Voids in Hardened Concrete*", ASTM International, 1997.
- [26] ASTM C 1585 - 04, "*Test Method for Measurement of Rate of Absorption of Water by Hydraulic-Cement Concretes*", ASTM International, 2004.
- [27] ASTM C 1012 - 07, "*Test Method for Length Change of Hydraulic-Cement Mortars Exposed to a Sulfate Solution*", ASTM International, 2007.
- [28] Shirayma, K. and Yoda, A., "*Proposed Methods of Test for Chemical Resistance of Concrete and Cement Paste in Aggressive Solutions*", Fifth International Conference on Durability of Building Materials and Components, 1991.
- [29] Van Tittelboom, K., et al., "*Test Methods for Resistance of Concrete to Sulfate Attack - A Critical Review*", Performance of Cement-Based Materials in Aggressive Aqueous Environments, vol. 10, Springer Netherlands, 2013, pp. 251–288.
- [30] JSTM C7401, "*Method of Test for Chemical Resistance of Concrete in Aggressive Solution*", Japanese Industrial Standard, 1999.
- [31] McCarthy, M. J. and Thomas, D. D., "*Pozzolanas and Pozzolanic Materials*" Kidlington, UK: Elsevier, 2019.
- [32] Khan, A.R. and Ganesh, A., "*The Effect of Coal Bottom Ash (CBA) on Mechanical and Durability Characteristics of Concrete*", Journal of Building Materials and Structures, vol. 3, no. 1, 1, 2016, pp. 31–42.
- [33] Stanish, K. D., et al., "*Testing the Chloride Penetration Resistance of Concrete: A Literature Review*", University of Toronto, 1997.



# DISTANCE AWARE TRANSMIT ANTENNA SELECTION FOR MASSIVE MIMO SYSTEMS

Shenko Chura<sup>1</sup>, Yalemzewd Negash<sup>1</sup> and Yihenew Wondie<sup>1</sup>

School of Electrical and Computer Engineering, Addis Ababa Institute of Technology, Addis Ababa University, Addis Ababa, Ethiopia

Corresponding Author's Email: [duskaanoo@gmail.com](mailto:duskaanoo@gmail.com)

## ABSTRACT

*Multiple-Input Multiple-Output (MIMO) antenna selection is a signal processing technique by which the Radio Frequency (R.F.) chain components are switched to their corresponding subset of antennas. Antenna selection resolves the complexity and power consumption of R.F. transceivers. This paper proposes an optimal antenna selection technique for multiple radio component type massive MIMO, which combines two selection techniques by exploiting the minimum signal-to-noise ratio (SNR) at the cell edge and dynamic channel condition due to mobility. After an adaptive selection has been made, the same number of R.F. components are active, and the rest are set to sleeping mode to apply fractional transmit power re-allocation at sub 6GHz and mm Wave frequencies. Accordingly, the branch with better signal quality among the array is chosen and added in iteration till the selected value is attained; however, re-selection still boosts E.E. at the cost of the total rate. The results show that the algorithm over performs the random selection, achieving better energy efficiency than full array utilization and random selection. Moreover, capacity reduction due to selection is compensated by applying nonlinear precoding at the cost of complexity.*

**Keywords:** Antenna Selection, Energy Efficiency, Massive MIMO, mm Wave, Precoding

## INTRODUCTION

Massive MIMO (mMIMO) is a large-scale MIMO device gaining popularity in wireless communications that scales up traditional MIMO by orders of magnitude, according to [1]. It takes into account multi-user MIMO, in which a base station with hundreds of thousands of antennas simultaneously supports many single-antenna terminals, as well as frequency resources. Connection reliability, spectrum quality, and radiated energy efficiency all improve when a device has many antenna elements. Each antenna element is linked to a single R.F. chain at the base station, which comprises mixers, analog-to-digital converters (ADC), and amplifiers [2]. Furthermore, the increase in the number of antennas and associated R.F. chains at the base station will result in physical restrictions, complexity, and expense [3]. According to [4], R.F. chains are responsible for approximately 50-80% of a base station's total transceiver power consumption.

As the number of quantization bits and B.S. antenna elements grow, the hardware complexity and power consumption of DACs grows exponentially. As a result, according to [5], adopting a low-resolution DAC is a potential choice. In addition,

conversion from analog to digital and digital to analog (ADC/DAC), phase shifters, and power amplifiers all affect the power amplifier's energy usage. Although the digital beam forming system provides a high data rate, the transceiver system uses the same number of antennas as the chains, resulting in excessive energy usage. On the other hand, a hybrid beamforming system utilizes fewer R.F. components and can provide equivalent spectral efficiency to a digital beamforming system while being more energy-efficient, according to [6].

Although hybrid beam forming is used to employ a small number of R.F. chains as the solution, one of the unanswered concerns is how to reduce some numbers among the whole array. Antenna selection has been employed as one of the power-saving approaches for a system with a vast array of R.F. components. The majority of recent studies in the literature have focused on performance analysis for large MIMO uplinks with low-resolution analog-to-digital converters. In this regard, [7] studied the impact of signal detection strategies on the energy efficiency of uplink MIMO systems with low-resolution analog-to-digital converters. In m MIMO systems, there have been few previous investigations on antenna selection. During the last few decades, various antenna selection strategies and algorithms have been investigated for classic MIMO.

The studies in [8] supported capacity-oriented selection criteria such as the greedy method and convex optimization. In [9], the authors introduced an antenna selection approach (AS) with a low degree of complexity that selects antennas with the least amount of constructive user interference. The suggested AS approach outperforms systems that use a more sophisticated channel inversion method when the transmitter uses precoders in

conjunction with a matched filter. The goal of the work in [10] was to eliminate the destructive fraction of the interference caused by the link between the sub streams of a modulated Phase Shift Keying (PSK) system. Singular value decomposition was used to offer a new Euclidean distance-dependent technique for antenna selection in spatial modulation systems that have a lower computing complexity than an exhaustive search [11]. Furthermore, the Symbol Error Rate (SER) approaches a complete search as the number of received antennas increases. As a result, the authors of [12] noted that, in comparison to previous and current research trends, there is still a lot of interest in mm Wave based massive MIMO antenna selection with less complexity, higher energy efficiency, and optimal data rates in recent years. This paper studies a system with transmit antenna selection for massive MIMO-enabled BS.

The technique is broken down into two sections: First, at the cell edge, a whole array device's energy efficiency (E.E.) is assessed using a fixed power allocation technique that assumes the channel is deterministic. In this scenario, the initial access condition is used to find the optimal number of B.S. antennas where the E.E. reaches its maximum. Second, as users go from the cell edge to the outskirts or center places, the minimum Signal to Noise Ratio (SNR) found at the cell edge is employed as a threshold value to further search for the ideal number.

To find is utilized as the total number of B.S. antennas. Then, the Free Space (F.S.) Path Loss (P.L.) model is employed for each mobile terminal, with adaptive power allocation based on minimum SNR at the cell edge. After determining, the subset of antennas with the best channel conditions is chosen, and E.E. is evaluated using spatial selectivity at mm Wave frequency ranges.



The following are the primary contributions of this paper:

Although our selection algorithm also shows that when partially applying the same exhaustive search technique, the complexity, and energy efficiency decrease and increase respectively due to double selection before exhaustive search.

- We derive a mathematical formulation for energy-efficient antenna selection.
- We evaluate the performance of mmWave based mMIMO antenna selection with and without nonlinear precoder from the perspective of energy efficiency.
- In comparison to prior methods, we analyze the computational complexity of the proposed transceiver system.

The rest of the work is structured as follows: A system model for mMIMO beam forming and array geometry is defined in section II. After the propagation model is explained in section III, antenna selection and power consumption models are followed in sections IV and V, where results and analysis are done. Finally, conclusions are drawn in section VI.

### SYSTEM DESCRIPTION

The system description is divided into two parts. First the signal model of a downlink massive MIMO system with transmit antenna selection is outlined. Then the addressed antenna selection technique is defined and formulated for a sub-optimal solution.

### System Model

Here, the channel model with only free space non-fading (i.e., pure) LoS propagation between the B.S. and the devices is presented; that the B.S. is fitted with a ULA with a  $\lambda/2$  spaced  $M$  antenna portion where  $\lambda$  is the signal wavelength where the presence of the mutual coupling effect between antenna elements is not taken in to account. There are randomly distributed  $K$  single-antenna devices inside the cell that simultaneously transmit data to the B.S. using the same time-frequency resources. Furthermore, it is also assumed that all of the  $K$  devices served by the B.S. are positioned at various angles on the far-field of the antenna array and undergo fading of large and small scales. Downlink  $M^o$  RF chains associated with  $K (K \leq M^o \leq M)$  are considered for a single-cell massive MIMO structure.

### Mobile Location Positioning

Mobile Location Positioning in today's cellular networks, identifying a mobile position is a critical problem. The angle of Arrival (AoA), Time of Arrival (ToA), and GPS are among the techniques used. In general, there are three methods for determining the location of a mobile terminal: satellite positioning, cellular network-based positioning, and indoor positioning. The trilateration method is used to calculate a mobile's location using a relative position of a base station (B.S.). Unlike the triangulation process, which requires the angle of each user for position tracking, only the distance between the B.S. and each user is required in this case [24].

### Close-In (CI) path loss model

The CI model is based on Friis and Bullington's fundamental radio propagation concepts, wherein the value is 2 for free space and 4 for the asymptotic two-ray ground bouncing model. It provides insight

into path loss as a function of the environment since base station towers are tall and inter-site distances for specific frequency bands are several kilometers. Previous UHF/microwave models employed a close-in standard distance of 1 km or 100m [12]. The CI 1m reference distance, as proposed in [13], is a suitable recommended norm that relates the real transmit power or PL to a reasonable distance of 1m. Standardization to a 1m reference distance simplifies dimension and model comparisons, provides a consistent description for Path Loss Exponent (PLE) and allows for quick and straightforward route loss estimates without the need for a calculator [14]. Using power control mechanisms, user terminals nearer to the BS are allocated lower power than those on the outskirts to control interference and fairness. CI path loss model is a generic frequency model that explains large-scale path loss at all applicable frequencies in a specific context. The dynamic range of signals perceived by users in a commercial system becomes significantly lower than the equation for the CI model, which is formulated as [12]

$$PL_{CI}(\cdot)_{dB} = PL_{FS}(f, 1m) + 10n \log_{10}(d) + \chi_a^{CI} \quad (1)$$

Where  $n = \sum(DA) / \sum(D^2)$ , denotes a single model criterion, the PLE, with 10n defining path loss in dB in terms large distance starting from 1m and  $(\cdot)$  represents frequency and distance parameters. The free space path loss,  $PL_{FS}(f, 1)_{dB}$  at 1m distance from a station and carrier frequency f.  $A = PL_{CI}(\cdot)_{dB} - PL_{FS}(f, 1m)$ ,  $10 \log_{10} d$  denotes a single model criterion, the PLE, with 10n defining path loss in dB in terms large distance starting from 1m and  $(\cdot)$  represents frequency and distance parameters. The free space path loss,  $PL_{FS}(f, 1)_{dB}$  at 1m distance from a station and carrier frequency f is given as

$$PL_{FS}(f, 1m)[dB] = 20 \log_{10}(4\pi/\lambda) \quad (2)$$

where  $\lambda$  is wavelength of the signal. It's worth noting that the CI model includes an intrinsic frequency interdependence of path loss in the 1m  $PL_{FS}$  value, and it only has one parameter compared to the ABG  $\alpha$ ,  $\beta$ , and  $\gamma$  model where  $\alpha$  and  $\gamma$  are coefficients showing the dependence

of path loss on distance and frequency, respectively and  $\beta$  is an optimized offset value for path loss in dB.  $\sigma^{CI} = \sqrt{\sum \chi_{\sigma}^{CI^2} / T}$  where T is the number of data points.

Table 1 shows the frequency ranges to be used in CI pathloss model in urban micro for street canyon (UMi-SC) and open space (UMi-OS) at line of sight (LOS) and non-line of sight (NLOS) conditions respectively [14]. As shown in the table, the CI model provides path loss exponent (PLE) of 2.0 and 1.9 in LOS, which approaches well with a free space PLE of 2.

Table 1: Close-in path loss model parameters

Scenario	Freq. (GHz)	Distance (m)	PL E/ $\alpha$	$\alpha^{CI}$
UMi-SC LOS	2-73.5	5-121	2.0	2.9
UMi-SC NLOS	2-73.5	19-272	3.1	8.0
UMi-OS LOS	2-60	5-88	1.9	4.7
UMi-OS NLOS	2-73.5	8-235	2.8	8.3
UMa LOS	2-73.5	58- 930	2.0	4.6
UMa NLOS	2-73.5	45-1429	2.7	10.0

### Trilateration Based Antenna Selection

In the selection process, the number of antennas to be selected is decided by adjusting the sufficient amount of transmit power to be radiated through only the

selected number of antennas. Trilateration is used to find a user's location so that the main beam can focus only on the desired location to minimize leakage. Due to user mobility, the transmit power adaptively changes as user position varies as a distance function. In this case, considering the minimum SNR at the cell edge as a threshold value, the number of transmit antennas can be reduced adaptively when the user comes closer to the center of the BS instead of using all arrays that may lead to unnecessary power wastage. In contrast, the BS only allocates the power proportional to the reduced distance to the maximum transmit power allocation concerning edge distance. Hence, only a few antennas are activated, as stated in (3). Then, the antennas with better channel gains are selected among the array using factorial permutations  ${}^M C_N$  as

$$M^o = \frac{M \cdot \sum_{i=1}^K p_t / K}{P_T} \quad (3)$$

where  $p_t$  is the transmit power adjusted for each user based on path loss,  $P_T$  is the total transmit power and  $N = M^o$  is the number of RF chain components. The selection process for the whole system is stated in a sub optimal algorithm 1 and 2 below.

**Algorithm 1:** Initial access based optimal number selection algorithm

**Input:**  $D_{max}$ ,  $M$ ,  $K$ ,  $f$ ,  $\gamma$ ,  $p_t$ ,  $B$ ,  $p_{amp}$ ,  $p_{bb}$ ,  $p_{syn}$ ,  $p_{dac}$ ,  $p_{mix}$ ,  $p_{filt}$

**Output:**  $EE$ ,  $M^*$ ; \*/

**begin**

```

1   $\zeta = 0$ ,  $\hat{p}_t = 30mW$ ,  $D_{max} = 300m$ ;
2  for  $l = 1 : \text{length}(M)$  do
3     $H \leftarrow (\text{randn}(K, l) + j(\text{randn}(K, l)))$ 
4     $\zeta(l) = \log_2(\text{real}(\det(I + (\frac{\gamma \text{sel}}{l}) H H^T)))$ 
5     $p_{tot(l)} \leftarrow p_{amp} + (p_{bb} + p_{syn}) + (l(p_{dac} + p_{mix} +$ 

```

$p_{filt}))$ ;

```

6   $EE \leftarrow \frac{\zeta(l)}{p_{tot(l)}}$ 
7    if  $l = M_{max}$  then
8       $M^* \leftarrow l$  ( $\text{find}(EE == \max(EE))$ )

```

In algorithm one, the parameters in each line are represented as follows:

- The outputs are energy efficiency and optimal number of antennas at cell edge respectively;
- In line 2,  $D_{max}$  is cell edge distance;
- $\zeta(l)$  in line 5 is capacity in each iteration;
- $p_{tot(l)}$  is total power in each iteration;

**Algorithm 2:** Number and element selection after reduced distance

**Input:**  $d_{min}$ ,  $\gamma$ ,  $p_t$

**Output:**  $EE$ ,  $M^o$

**begin**

```

1  re-trilateration: for  $i \in k$  do
2     $r_k \leftarrow R(k_i)$ 
3    if  $r_k \neq d_{min}$  then
4       $P_{rmin} \leftarrow P_{tmax} / \Gamma(R)$ 
5       $P_r(k) \leftarrow \Gamma_r P_{rmin}$ 
6       $M_1^o = \frac{M \cdot \sum_{i=1}^k P_t / K}{P_T}$ 
7       $M_2^o \leftarrow \frac{\text{round}(\sum(\Gamma(r)) / K) M}{\Gamma(R)}$ 
8      if  $M_1^o \neq M_2^o$  then
9        goto re-trilateration
10      $M_1^o \leftarrow M_2^o$ 
11     for  $\gamma = 1 : M^o$  do
12        $\Psi = \text{rand}(K; M) + j\text{rand}(K, M)$ 
13        $H = \frac{\Psi}{\sqrt{M^o}}$ ;
14       for  $M_i^o = 1 : M^o$  do
15          $H_c = [H ; [M_i^o \quad M - i]]$ 
16          $\Phi = \det(I + \gamma * H_c * H_c^T)$ 
17          $\zeta(m) = \log_2(\text{real}(\Phi))$ 
18          $\zeta_{max} = \max(\zeta(m))$ 

```

18  $M_i^0 \leftarrow \text{find}(\zeta = \zeta_{\max})$   
 19  $\zeta(\gamma) \leftarrow \zeta_{\max}$   
 20  $EE \leftarrow \frac{\zeta(\gamma)}{p_{\text{tot}}}$  ;

In algorithm one, the parameters in each line are represented as follows:

- The outputs are minimum distance, fixed SNR and total transmit power respectively;
- $r_k$  in line 3 is the random distance of a user;
- $\Gamma(R), \Gamma(r)$  in line 8 is path loss in dB at cell edge and reduced distance;
- $Pr_{\min}$  in line 5 is minimum received power at the cell edge;
- $P_r(k)$  in line 6 is received power at a k user;
- EE in line 22 energy efficiency with selected branches and total power;
- $M_1^0$  in line 7 is the number of optimal antennas at reduced distance;

### Energy Efficiency Evaluation

With chosen antennas  $M^o$ , among M, the transceivers corresponding to  $M^o$  are turned on while some M - $M^o$  's shut off. With massive MIMO, the number of BS antennas (M) is assumed to be always much greater than the number of single antenna user terminals ( $M \gg K$ ) and allow  $M^o$  to be within the range from K to M. Where  $M^o$ , K is the number of antennas to be chosen and the total number of user terminals with a single antenna respectively. The downlink-channel model is

$$y_l = \sqrt{\rho K} H_l^{(M^o)} z_l + n_l. \quad (4)$$

Where  $H_l^{M^o}$  is a  $K \times M^o$  channel matrix on carrier  $l$  and the  $M^o$  subscript indicates that antenna selection has been made, i.e.,  $M^o$  columns of  $H_l^{M^o}$  are chosen from the complete channel matrix of  $K \times M$ .

### Dirty Paper Coding Sum Capacity ( $C_{(DPC)}$ )

The downlink sum-capacity at subcarrier is given by [6]:

$$C_{DPC_l} = \max_{P_l} \log_2 \det (I + \rho K (H_l^{(N)})^H P_l H_l^{(N)}) \quad (5)$$

In (5),  $P_l$  is a diagonal power allocation matrix with  $P_{l,i}$   $i = 1, 2, \dots, K$  on its diagonal. And the optimization is also carried out according to the total power restriction of  $\sum_{i=1}^K P_{l,i} = 1$  as in (5). This problem of optimization is convex, and can be solved, for example, by using the water-filling algorithm of sum-power iterative. DPC is highly complex to implement in practice. However, there are suboptimal linear precoding schemes, such as zero-forcing (ZF) precoding that is much less complex and performs fairly well for massive MIMO [15].

### Zero Forcing Sum Capacity ( $C_{\{ZF\}}$ )

The total rate achieved by ZFT is [15]

$$C_{ZF,l} = \max_{Q_l} \sum_{i=1}^K \log_2 (1 + \rho K Q_{l,i}) \quad (6)$$

Where  $Q_{l,i}$  represents SNRs obtained by the various users and the maximization is carried out according to the total power constraint

$$\sum_{i=1}^K Q_{l,i} \left[ \left( H_l^{(N)} (H_l^{(N)})^H \right)^{-1} \right] = 1 \quad (7)$$

In (6) and (7),  $Q_l$  is a diagonal matrix with  $Q_{l,i}$   $i = 1, 2, \dots, K$  in its diagonal, and  $[.]_i$  means the matrices  $I$  diagonal. The  $\left( H_l^{(N)} (H_l^{(N)})^H \right)^{-1}$  diagonal elements reflect the power penalty of null-out intervention.

An  $M \times M$  diagonal matrix of  $\varphi$  with binary diagonal elements has been implemented to

pick the  $N$  columns from the complete MIMO matrix  $H_l$ .

$$\phi_i = \begin{cases} 1 & \text{Selected} \\ 0 & \text{Otherwise} \end{cases} \quad (8)$$

indicating whether the  $i^{th}$  antenna is selected, and satisfying  $\sum_{i=1}^M \phi_i = N = M^o$ . Using Sylvester's determinant identity,  $\det(I+AB) = \det(I+BA)$ , the DPC sum capacity in (5) can be re-written in terms of  $\phi$  as

$$C_{DPC,l} = \max_{\phi_l} \log_2 \det(I + \rho K P_l H_l^{(N)} (H_l^{(N)})^H) = \max_{\phi_l} \log_2 \det(I + \rho K P_l H_l \phi(H_l)^H) \quad (9)$$

subject to  $\sum_{i=1}^K P_{l,i} = 1$ .

The optimal  $\phi$  (common to all subcarriers) is found by maximizing the average DPC capacity,

$$\phi_{opt} = \max_{\phi} \frac{1}{L} \sum_{l=1}^L \log_2 \det(I + \rho K P_l H_l \phi(H_l)^H) \quad (10)$$

With the subsequent range of antenna, the respective sum-rate of ZF

$$C_{ZF,l} = \max_{Q_l} \sum_{i=1}^K \log_2(1 + \rho K Q_{l,i}) \quad (11)$$

$$\text{Subject to } \sum_{i=1}^K Q_{l,i} \left[ (H_l^{(N)} (H_l^{(N)})^H)^{-1} \right]_{i,i} = 1 \quad (12)$$

Despite  $\phi_{opt}$  may not be optimal for ZF, the ZF sum-rate indicates the antenna selection performance when using a more practical precoding scheme than DPC. As discussed above, exhaustive search of all possible combinations of  $N$  antennas will certainly give us the optimal  $\phi$  however, it is extremely complex and infeasible for massive MIMO. From (11, 12), it can be seen that zeroing the upper and lower matrix elements requires additional power consumption in ZF however still it is simpler in processing compared to DPC which has no additional power penalty and complex on the other hand.

## Energy Efficiency Evaluation

The total energy efficiency of the system can be evaluated as [15]:

$$C = KE [\log_2 ([1 + \rho |g_k|^2])] \quad (13)$$

$$EE = KE [\log_2 (1 + \rho |g_k|^2)] / P_{total} \quad (14)$$

Where  $P_{total} = P_{amp} + P_{CP}$  and  $P_{CP} = P_{bb} + P_{syn} + M^o * (P_{dac} + P_{mix} + P_{filt})$  where  $P_{CP}$  accounts for the circuit power consumption.  $P_{amp}$  is the amount of the power produced by various analog components. In  $P_{CP}$ , baseband signal processing ( $P_{bb}$ ), synchronization ( $P_{syn}$ ) are independent of number of BS antennas while digital to analogue conversion power ( $P_{dac}$ ), mixing ( $P_{mix}$ ) and filtering ( $P_{filt}$ ) power linearly increase with selected BS antennas. Table 2 contains the parameters to be used for simulation purpose in evaluation of EE according to (14).

Table 2: Complexity Analysis

Algorithm 2	Algorithm 1+2
$\left( \hat{n} \binom{M}{M^o} \right)$	$\left( \hat{n} \binom{M}{M^o - 1^*} \right)$

The table states the combinational permutation of the algorithms which we compare with that of [16], [17], [18] and [19] which accounts for  $\hat{n} \binom{M}{M^o}$  where  $\hat{n} = M^2 + 2M_{o \equiv s}^2 + M^o$  and  $1^*$  is the deducted elements due to selection. According to [17] and [18], the computational complexity due to selection process is shown in (16) and (17) respectively.

$$\mathcal{O}_1(.) = 16n^3 + \hat{n}^2 (24M^2 + 40M + 24 - 24M^{o^2} - 24M^o), \quad (15)$$

$$\mathcal{O}_2(.) = \hat{e} + 20(M^2 + M - M^{o^2} - M^o), \quad (16)$$

$$\mathcal{O}_3(.) = \mathcal{O}_1(.) + \mathcal{O}_2(.), \quad (17)$$



$$\mathcal{O}_4(.) = \hat{n}(M^{o^2}(M3(M+1)))(18)$$

Where  $\hat{e} = \hat{n}(34M^2 + 44M - 36M^{o^2} - 34M^o)$  and  $(.)$  denotes  $(M, M^o)$ . In [30] low-complexity transmit antenna selection (LCTAS) was studied and found to have complexity level as follows:

$$\mathcal{O}_{[30]} = \mathcal{O}(M^o S^o \binom{M}{M^o})(19)$$

where  $S^o$  is the number of symbols for a constellation type.

Table 3: Simulation parameters for the overall work

Parameter	Value	Description
$P_{tx}$	5mW	Transmission power
$p_{mix}$	0.033	Mixing consumption
$p_{filt}$	0.02	Filtering consumption
$p_{bb}$	0.03	Base band signal processing power
$p_{syn}$	0.05	Synchronization power
$p_{dac}$	0.015	Digital to analogue conversion power
$p_{amp} = p_{tx}/\eta$	$\eta=0.01$	Amplifier power
$f_s$	1800 Mhz	Sub 6Ghz frequency
$f_m$	37 Ghz	mmWave band

## RESULTS AND DISCUSSIONS

In figure 1, ergodic capacity of different MIMO configurations for iid (independent identically distributed) channel has been shown. From the figure it can be concluded that, the capacity becomes higher for of massive MIMO with different number of antenna configurations than classical MIMO systems however at lower SNR level the difference is much less and can be neglected. However, increase in the number of BS antennas accounts for increase in SNR

which signifies system's capacity enhancement. In this case, NT represents the number of BS antennas, M.

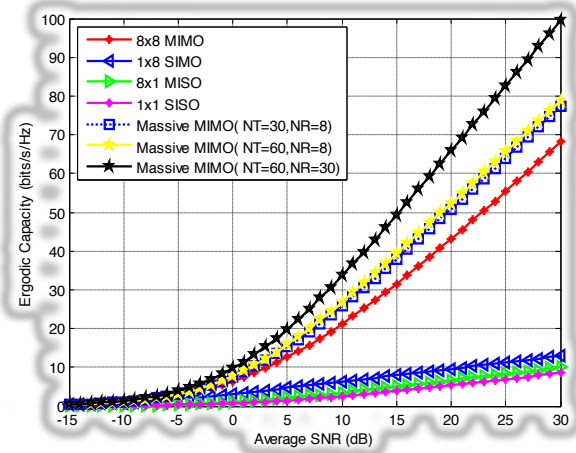


Figure 1: Ergodic Capacity for i.i.d. Rayleigh fast fading channel in different MIMO configurations.

Figure 2 depicts the effect of randomly selected transmit antennas on system energy efficiency. The energy efficiency increases with the number of transmit antennas (M) first, and after an optimal point, it abruptly declines. The increase in BS antennas is directly associated with the rise in the corresponding radio frequency chain components, which accounts for enormous power consumption in a system. From the figure, the optimal number of antennas ( $M^*$ ) also depends on the number of user terminals (K). For K=5, 10,15, and 20,  $M^*=5, 7, 8$ , and 9.

This turning point is when the system's total power consumption exceeds the increase in the full rate. Hence, the number of antennas to be selected should not exceed this point to maintain EE; however, finding the optimal point also has its challenge due to processing complexity.



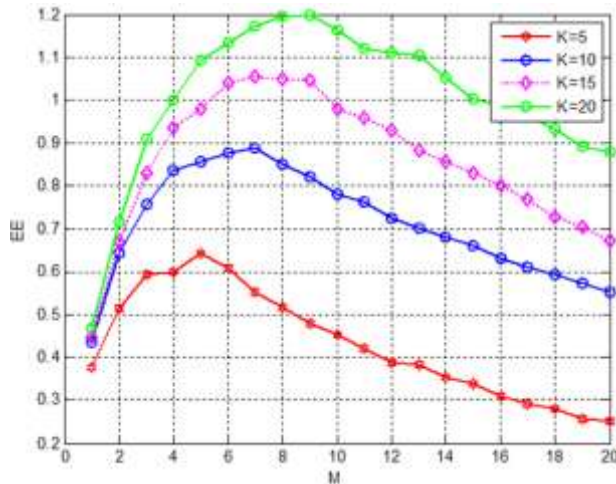


Figure 2: EE at different users and BS antenna settings.

The correlation between energy efficiency,  $k$ , and  $M$  in a massive MIMO system with statistical and instantaneous SNR values is depicted in figure 3. The outcome is evaluated for cell-edge users in LoS settings utilizing algorithm one processes. According to the result, while energy efficiency initially rises as  $M$  increases, it begins to drop at some point as  $M$  keeps growing. For the same  $k$ , statistical and instantaneous or fixed SNR are compared in this figure. Accordingly, fixed SNR outperforms for small  $M$  and comes up short for large  $M$ . It has also been proven that EE grows with user terminals. The EE values for  $k=20$  are obtained from the average value of both statistical and fixed SNR values.

On the other hand, EE presents multiple optimal points due to unpredictable channel circumstances. Furthermore, depending on the number of users and SNR modalities, the ideal EE point for each arrangement differs.

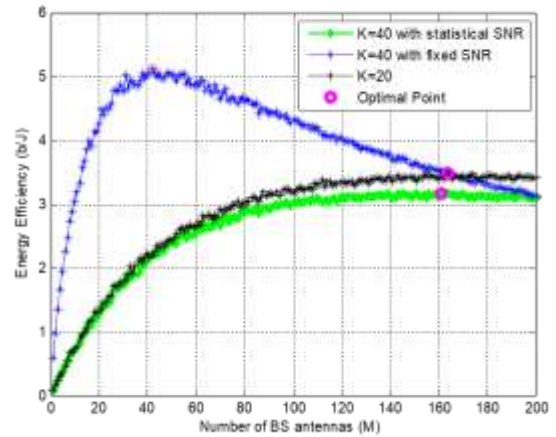


Figure 3: EE for statistical and fixed SNR at sub 6Ghz.

In figure 4, the effect of channel variation on total power and the optimal number of antennas to be selected is shown. When statistical channel variation is considered, the SNR varies. Therefore both total power and  $M^*$  grow large to combat small scale fading by adaptively allocating the desired amount of power. With fixed SNR, fewer antennas can achieve an optimal level than statistical SNR. From the figure, evaluation with statistical SNR accounts for total power consumption than instantaneous SNR assumption, which is 20mW and nearly 19mW for statistical and fixed SNR, respectively.

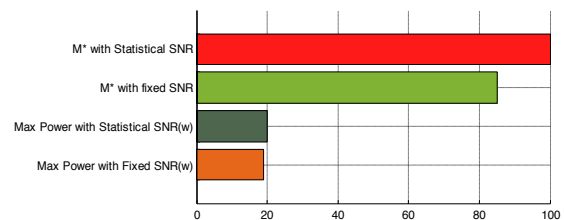


Figure 4: Optimal number of antennas and maximum power for statistical and fixed SNR.

Figures 5 and 6 present the results according to the proposed algorithm by combining the three scenarios, and we have compared the performance of each at CI and FSPL using mmWave and sub 6GHz frequency ranges. The first scenario is finding  $M^{\star}$  from the full array at the indoor cell edge, finding according to (3), and finally evaluating capacity values by combinational permutation as  $\binom{M}{M^o}$ . At the initial access, we assume a deterministic channel, equal power allocation among all BS antenna elements and the point at which the EE graph starts diminishing is evaluated using the reference signal. Then the number of antennas at that point is used as a baseline for our further considerations. Before the energy efficiency evaluation process, we make the analysis of free space and CI path loss models according to their formulations stated in (1) and (2).

Accordingly, the FS model provides higher data rates due to obstruction; however, CI is more realistic than FS in practical scenarios. Based on this intuition, we have applied an antenna selection algorithm for both, and the results show that a much smaller number of antennas are selected in free space than CI. Besides, when CI path loss is applied to mmWave and sub 6GHz frequency ranges and reached for fixed total system power, CI with sub 6GHz is more energy-efficient than mmWave. Despite high-frequency signals carrying larger data than low-frequency signals, as frequency increases, the blockage due to different impairments also exhibits low wavelength, which negatively affects the received signal.

Low received signal accounts for low data rate at the receiver, and thus EE is degraded compared to CI. Finally, we have found that the FS path loss with the DPC precoder changes the graph from logarithmic to almost linear and starts an abrupt shift to decline after the maximum point. However, it is limited to the total value in this case.

Figure 5 depicts minimum SNR-based antenna selection using linear and nonlinear precoders and compares with EE at full array implementation with no precoders. After finding an optimal number of antennas, as figure 5, it applies (3) to recalculate a new optimal point that depends on the users' current position or distance and adaptive reduction of  $M$  instead of transmitting power. In this case, the optimal, which was found in full array implementation, is used as  $M$  to re-search the new optimal value (3). Despite the reduction in the total rate when the number of antennas is reduced, the reduction in total power consumption compensates for maintaining EE. Finally, applying precoders in general and nonlinear DPC, in particular, boosts the total rate of the system and EE as well. We have evaluated EE as a function of BS antennas at different power levels for full array and selection implementations. The performance of the system has also been assessed with and without the nonlinear precoding and shown that antenna selection with minimum SNR significantly improves the energy efficiency with less transmit power and DPC precoder.

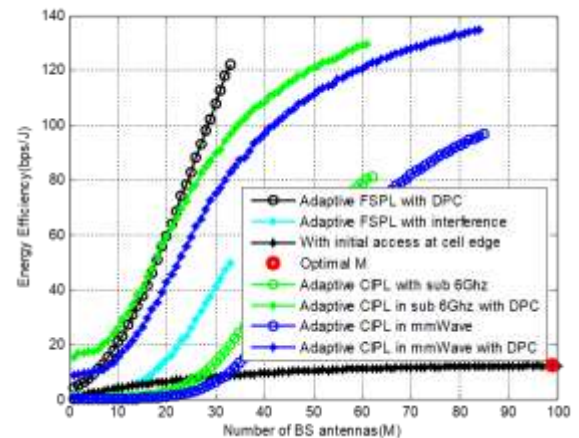


Figure 5: Energy efficiency evaluation as a function of number of BS antennas with at mmWave frequency,  $f=38$  GHz  $M=64$ .

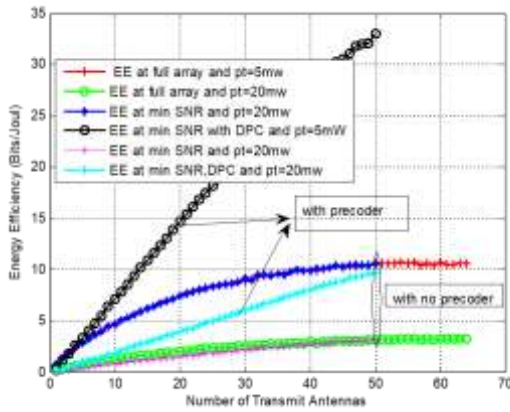


Figure 6: Energy efficiency evaluation as a function of number of BS antennas with at mmWave frequency,  $f=38$  GHz  $M=64$ .

The complex nature of the proposed selection algorithm is shown in figure 7, and it is compared with the works that employ comparable strategies. The number of iterations of the main and nested loops that must occur when selecting the branch with the best channel gain among the complete array is referred to as complexity in this scenario. On the other hand, random selection is minor complex, despite having a lesser capacity than complex selection, as shown in the graph. This is because the selection is made regardless of channel gain, which is critical in increasing capacity and complexity.

For random selection, the number of iterations to select  $M$  antennas is only one as it has no combination with the channel branches. Our proposed algorithm is also compared with [16], [17], and [18], which are among the simplest and follow similar approaches to the best of our knowledge. The complexity order of each is our proposed technique and random selection according to (17), (18), and (19). We have also found that the proposed algorithm is more energy-efficient than random at the cost of some complexity which is less than that of [16] and [19]. Moreover, the energy efficiency of the proposed technique has

been shown to surpass random selection, full array utilization, and some other literature, as shown in the figure. However, the effect and trade-off rate, including EE of the literature above, is left as our future work.

Therefore, the selection technique meets our primary goal of proposing an energy-efficient system at manageable complexity.

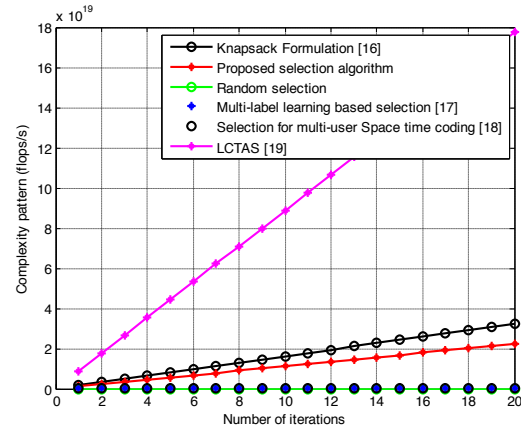


Figure 7: Computational complexity of selection algorithms with adaptively selected number of elements and  $M=64$ .

## CONCLUSIONS

This work has focused on the problem of system energy efficiency due to the massive number of antenna elements to be installed on a single BS in the upcoming wireless communication era. Adaptive antenna selection technique has been proposed as a novel strategy in resolving a substantial amount of power consumption and complexity as a result of power-hungry RF elements which grow with antenna elements. The selection has been made for cell-edge users with a full array at fixed power allocation and minimum SNR-based selection for cell center users. Both cases are used to achieve the optimal number of antennas at which EE becomes maximum. The key idea of the proposed algorithm is to minimize the number of RF chains, and performance evaluation has been done in

several scenarios by applying precoders at different frequency ranges. The numerical results show that the proposed antenna selection algorithm performs better than full utilization of the array while finding some computational complexity when applying nonlinear precoders to compensate the total rate whilst selection gets negative effects. Moreover, we have evaluated the complex pattern of previous and current works with similar techniques. Accordingly, it has been shown that the proposed approach is least complicated and energy-efficient compared to the Knapsack formulation and LCTAS.

## REFERENCES

- [1] Larsson E., Tufvesson G., F., Edfors O., and Marzetta T. L., "Massive MIMO for next generation wireless systems," *CoRR*, vol. abs/1304.6690, 2013.
- [2] Heath R.W., Sandhu S., and Paulraj A., Antenna selection for spatial multiplexing systems with linear receivers, *IEEE Commun. Lett.*, vol. 5, no. 4, pp. 142-144, Apr. 2001.
- [3] Correia L., Zeller D., Blume O., Ferling D., Jading Y., Godor L., Auer G., and Van der Perre L., Challenges and enabling technologies for energy aware mobile radio networks, *IEEE Communications Magazine*, vol. 48, no. 11, pp. 6672, November 2010.
- [4] Heath R.W., Sandhu S., and Paulraj A., Antenna selection for spatial multiplexing systems with linear receivers, *IEEE Commun. Lett.*, vol. 5, no. 4, pp. 142-144, Apr. 2001.
- [5] Gharavi-Alkhansari M., and Gershman A.B., "Fast Antenna Subset Selection in MIMO Systems," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 339-347, Feb. 2004.
- [6] Dua A., Medepalli K., and Paulraj A.J., "Receive Antenna Selection in MIMO Systems Using Convex Optimization," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2353-2357, Sept. 2006.
- [7] Wang B.H., Hui H.T., and Leong M.S., Global and Fast Receiver Antenna Selection for MIMO Systems, *IEEE Trans. Commun.*, vol. 58, no. 9, pp. 2505-2510, Sept. 2006.
- [8] Xu Z., Sfar S., and Blum R. S., "Analysis of MIMO systems with receive antenna selection in spatially correlated Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 251-262, Jan. 2009.
- [9] Masouros C. and Alsusa E., "Dynamic linear precoding for the exploitation of known interference in MIMO broadcast systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1396-1404, March 2009.
- [10] Gesbert R., "Soft linear precoding for the downlink of DS/CDMA communication systems," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 1, pp. 203-215, January 2010.
- [11] Xiang G.; Edfors, Ove; Liu, Jianan; Tufvesson, Fredrik, "Antenna selection in measured massive MIMO channels using convex optimization," *IEEE GLOBECOM Workshop*, 2013, Atlanta, Georgia, United States.
- [12] Gao X., Edfors O., Rusek F., and Tufvesson F., "Massive MIMO performance evaluation based on measured propagation data," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3899-3911, July 2015.

- [13] Rappaport T. S., *Wireless Communications: Principles and Practice*, 2nd ed," *Upper Saddle River*, NJ: Prentice Hall, 2002.
- [14] Rappaport T. S. et al., "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design (Invited Paper), *IEEE Transactions on Communications*," vol. 63, no. 9, pp. 3029-3056, Sep. 2015.
- [15] Guthy C., Utschick W., and Honig M., "Large system analysis of sum capacity in the gaussian MIMO broadcast channel," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 149-159, Feb. 2013.
- [16] Husbands R., Ahmed Q., and Wang J., "Transmit Antenna Selection for Massive MIMO: A Knapsack Problem Formulation," *IEEE ICC 2017 Wireless Communications Symposium*, Jan, 2017.
- [17] Yu W., Wang T. and Wang S., "Multi-Label Learning Based Antenna Selection in Massive MIMO Systems," DOI 10.1109/TVT.2021.3087132, *IEEE Transactions on Vehicular Technology*, June 09, 2021.
- [18] Kim S., "Efficient Transmit Antenna Subset Selection for Multiuser Space-Time Line Code Systems," *Sensors* 2021, 21, 2690. <https://doi.org/10.3390/s21082690>.
- [19] Pillay N., Xu H. , "Low-complexity transmit antenna selection schemes for spatial modulation," *IET Commun.*, 2015, Vol. 9, Iss. 2, pp. 239–248 doi: 10.1049/iet-com.2014.0650.



# ADDIS ABABA LIGHT RAIL TRANSIT SYSTEM ENERGY FLOW ANALYSIS

Asegid Belay<sup>1</sup> and Getachew Biru<sup>2</sup>

<sup>1</sup> African Railway Center of Excellence, Addis Ababa Institute of Technology, AAU

<sup>2</sup> School of Electrical & Computer Engineering, Addis Ababa Institute of Technology, AAU  
Corresponding Author's Email: Assegidjesus@gmail.com

## ABSTRACT

*With the continued focus on growing energy prices and environmental concerns, lowering energy consumption and maintaining the environmental sustainability of railway systems is becoming a crucial problem to which greater attention is being paid. In recent years, urban rail systems have grown in popularity as a method for reducing traffic congestion and pollution in metropolitan areas. Despite the fact that the railway system is likely the most energy-efficient mode of land-based transportation, there is still potential for improvement. In this regard, significant amounts of energy can be saved by installing energy storage on an electrified transit system allowing energy from braking to be captured. However, the amount of energy saved is dependent on the amount of energy transferred during braking, which relies on the drive cycle and the vehicle parameters. The overall benefit can be determined by analyzing the energy flow through components in an electrified transit system. In this paper, electrified transit system energy flows are analyzed for Addis Ababa light rail transit system. The methodology used assesses energy flows in the traction system, establishing where energy is dissipated. The analysis is performed for a specified drive cycle. Finally, the analysis showed that 37.9 % of the total energy loss over a drive cycle could be saved in Addis Ababa light rail transit system.*

**Keywords:** braking energy, energy flow, energy efficiency, environmental sustainability, light rail transit,

## INTRODUCTION

Today, there is a growing emphasis on the environmental consequences of all government initiatives and policies. Sustainability is becoming increasingly

important as people have a better understanding of the reactive changes and deterioration of planet Earth that we are experiencing on daily basis. As Rohit Sharma and Peter Newman sustainability is defined as "development that fulfills current demands without jeopardizing future generations' ability to meet their own needs." Ultimately, the sustainable development method is the only way that humanity can pursue for future generations. Without it, the earth's natural resources will be gone, leaving only synthetic materials. These options involve a system of boundaries in time (25–50 years), space (micro and macro-levels), and domain (social, economic, and environment) [2].

Sustainability is also an important factor in the construction of rail transportation. As governments look to the future for sustainable transportation, electrified rail networks have given and will continue to provide a mode that uses renewable energy. According to studies, rail is naturally more efficient than road transport, and when combined with renewable energy, it may provide a long-term source of mobility for future generations while reducing emissions [3]. Today, it is apparent that urban rail systems play an important role in the sustainable development of metropolitan cities like Addis Ababa for a variety of reasons, the most important of which is their relatively low energy consumption to transit capacity ratio. Nonetheless, major improvements in energy efficiency are required to maintain their environmental benefits over other modes of transportation in a context defined by increasing capacity demands and energy prices. There is widespread consensus that railways have significant energy savings potential, both short and long terms. Whereas technology advancements in rail cars will be gradual and take time to diffuse, there are



numerous viable short and medium-term saving techniques targeted at optimal control and the utilization of current technologies or operational enhancements. Consequently, energy efficiency has emerged as a prominent subject within railway industry. For example, 28 European railway Operators have committed to reduce CO<sub>2</sub> emissions per passenger kilometer and per ton kilometer by 50% by 2030 [4].

In addition to improvements in vehicle technology, infrastructure and building design, and loading of freight trains, the European union commissioned study [5] identifies several potential areas, which can achieve efficiency improvements in railway systems such as weight reduction, reduction of air resistance, optimization of space utilization, improvements in electric traction components, efficiency gains in diesel traction technology, recovery of braking energy, reduction in energy consumption of comfort functions, energy efficient driving, traffic flow management, improvement of occupancy rates, energy meters and management and organization.

All these areas can be investigated to achieve energy efficiency improvements. It is also essential to consider the interactions between the different areas, particularly when efficiency gains are achieved by installing components, which increase the weight of electric vehicles and affect space utilization. On the other hand, recently energy flow assessment has gained huge momentum in railway system. Research undertaken in [6] performed a comprehensive analysis to determine where energy is dissipated in traction system.

The result consists of an analysis of energy movements into and out of trains as shown in Fig.1, taking into account losses. This research paper developed methods used to analyses energy flows in Addis Ababa light rail electrified transit system. The analysis is used to assess the energy dissipated through braking, hence establishing the amount of energy that could be potentially saved.

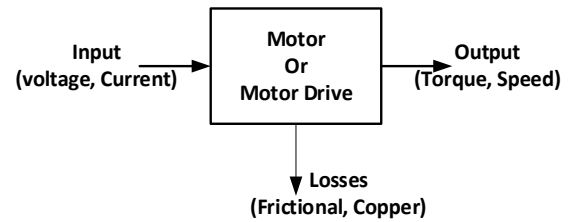


Fig.1 Energy flow block diagram

### Energy flow in light rail transit

An electric transit vehicle converts electrical energy into kinetic and potential energy. Energy is dissipated overcoming frictional forces and in braking. Energy is also dissipated through other mechanisms, including driveline losses and to power auxiliary loads. The electrical energy is transmitted from a local distribution network, through a traction substation, and the traction supply system. These stages also have energy losses. Fig.2 shows a more comprehensive representation of energy flows in an electrified transit system, including loss mechanisms. This paper considers the energy flow through each sub system identified in Fig.2 determine the overall energy dissipation.

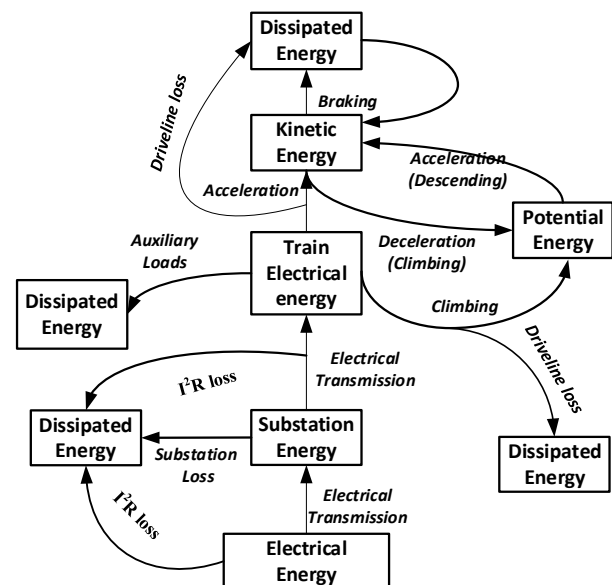


Fig.2 Energy flow in railway system

### Tractive resistance

In all traction applications, energy is consumed overcoming frictional forces. Davis [7] describes tractive resistance,  $F_r$  as a quadratic function of the speed of a vehicle.

$$F_R = a + b \left( \frac{ds}{dt} \right) + c \left( \frac{ds}{dt} \right)^2 \quad (1)$$

The first term  $a$ , is a constant with respect to speed (except at zero speed) but varies with the mass of the vehicle. The constant term is made up of two components, journal resistance and the static component of rolling resistance. The second term,  $ds/dt$  includes resistive forces which are proportional to speed, including the dynamic component of rolling resistance. The third term  $(ds/dt)^2$  represents aerodynamic drag forces. Curvature resistance adds another element to tractive resistance, however this can be considered negligible. The coefficients of the Davis equation can be calculated by considering the laws of physics, taking into account aerodynamic drag, rolling resistance and static friction. The parameters are often determined experimentally [8], fitting the coefficients to curves obtained through run-down tests. If the velocity of a vehicle is considered as a function of time,  $ds(t)/dt$ , the resistive force can be described as a function of time  $F_R(t)$ ,

$$F_R(t) = a + b \left( \frac{ds(t)}{dt} \right) + c \left( \frac{ds(t)}{dt} \right)^2 \quad (2)$$

The Power,  $P_R(t)$  dissipated by the train to overcome a frictional force is the product of force and speed,

$$P_R(t) = F_R(t) \cdot \frac{ds(t)}{dt} \quad (3)$$

Then using (2) and (3)

$$P_R(t) = a \frac{ds(t)}{dt} + b \left( \frac{ds(t)}{dt} \right)^2 + c \left( \frac{ds(t)}{dt} \right)^3 \quad (4)$$

This can be integrated with respect to time to determine the energy losses caused by frictional forces.

$$E_R(t) = \int_{t_0}^t a \frac{ds(t)}{dt} + b \left( \frac{ds(t)}{dt} \right)^2 + c \left( \frac{ds(t)}{dt} \right)^3 \quad (5)$$

## The driveline

The typical energy flow diagram via the DC-fed railway is illustrated in Fig. 2 to evaluate the overall energy efficiency of the system from the substation to the train. In terms of levels, there are three layers: substation level, catenary system level, and train level. The substations take electricity from the national power grid to power the whole railway system. After losses from substations, the remaining substation energy can be transmitted to the catenary. Some energy is wasted as heat when the current passes through the resistive catenary wire. The resistance of the transmission conductor is a time-varying parameter determined by the position of the trains and the network. The train receives energy through its pantograph. Finally, energy reaches to the traction motors through mechanical transmission system and power electronics converter as shown in Fig.3.

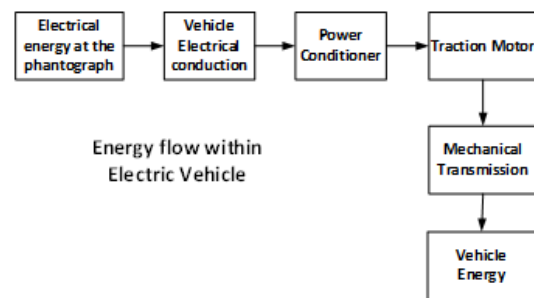


Figure 3 Electric train Energy flow block

## Train electrical conductors

Electric energy is transported by electric conductors from the pantograph to the electronic power converter, which dissipates heat energy. The cable lengths in the train are usually short and the resistance levels are low, which can be considered as insignificant transmission losses inside the vehicle.

## Power conditioner

In order to achieve the necessary torque, the supply of electrical power is controlled by means of power electronic converters. Many new train use induction motors powered by inverter. Inverters transform a DC supply to the VVVF (three-phase variable voltage frequency) supply. Modern inverter consists of six switches (IGBTs with anti-parallel diodes) as shown in Fig.4. The switches are working

to provide the appropriate frequency and voltage for a three-phase AC output.

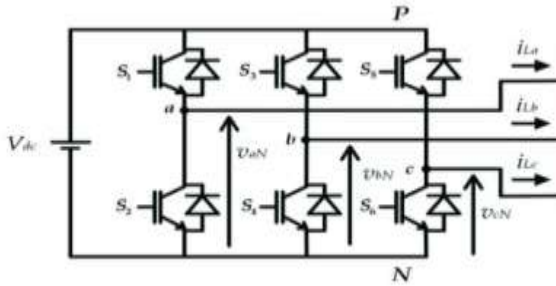


Fig.4 Inverter Circuit Diagram

The loss mechanisms of an inverter can be divided into two categories: conduction and switching [9]. Equation (6) describes the losses in one leg of an inverter  $P_{LLS}$

$$P_{LLS} = P_{cD} + P_{cQ} + P_{s,onQ} + P_{s,offQ} \quad (6)$$

where  $P_{cD}$ =conduction losses in the diodes

$P_{cQ}$  = conduction losses in the IGBTs

$P_{s,onQ}$  = switch-on losses in the IGBTs

$P_{s,offQ}$  = switch-off losses in the IGBTs

The diode conduction,  $P_{cD}$  losses can be described as

$$P_{cD} = \frac{1}{T} \int_0^T V_{\alpha}(t) i_{\beta}(t) dt \quad (7)$$

where  $V_{\alpha}(t)$  is the diode junction voltage and  $i_{\beta}(t)$  is the diode current, when the diode conducts this equals the line motor current.  $V_{\alpha}(t)$  can be written as a function of the current as shown below,

$$V_{\alpha}(t) = V_{ao}(t) + i(t)R_{\delta} \quad (8)$$

The values for  $V_{ao}(t)$  and  $R_{\delta}$  can be derived [9] from manufacturer's datasheets [10]. For a Siemens BSM75GB120 these values are  $V_{ao}(t) = 1.25V$  and  $R_{\delta} = 7.71m\Omega$ . The IGBT conduction losses can be described as

$$P_{cQ} = \frac{1}{T} \int_0^T V_{\psi}(t) i_{\psi}(t) dt \quad (9)$$

where,  $V_{\psi}(t)$  is the IGBT junction voltage and  $i_{\psi}(t)$  is the IGBT forward current, which equals the motor current, when the IGBT is

conducting,  $V_{\psi}(t)$  varies is a function of the current.

$$V_{\psi}(t) = V_{\psi o}(t) + i_{\psi}(t)R_{\psi} \quad (10)$$

where  $V_{\psi o}$  is 1.83V and  $R_{\psi}$  is 17.33m $\Omega$  [10].

The switching losses,  $\gamma_{on}$  and  $\gamma_{off}$  are a function of the IGBT forward current. Fig.5 shows the switching losses for a Siemens device, BSM75GB120.

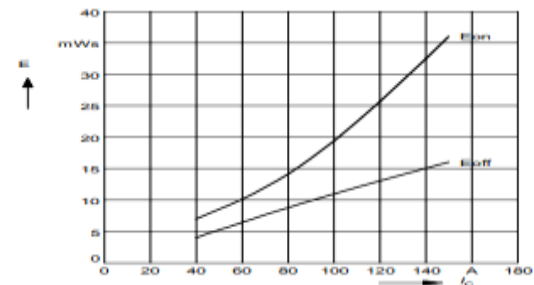


Fig.5 Switching losses against current for a Siemens BSM75GB120 device [10]

The energy loss over one cycle is the sum of the energy losses from each switching operation. The switching losses vary over the sine wave. To simplify the calculation, the RMS current is used to find the switch on and switch off losses from the graph. The switching power ( $P_s$ ) can be described as.

$$P_s \approx \epsilon_{\phi}(\gamma_{on} + \gamma_{off}) \quad (11)$$

The inverter loss energy is dissipated as heat, and therefore there is a cooling requirement. For the purpose of analyzing energy flows, the gate drive control circuit and cooling energy requirements are considered as auxiliary loads.

### The induction motor

Induction motor loss mechanisms include ohmic losses, iron losses and frictional and wind age losses [11]. The frictional and wind age losses can be considered as part of the train's frictional loss. When the coefficients of the Davis equation are determined, the motor frictional and wind age losses are included in the tractive resistance analysis. The ohmic losses that occur in the stator and rotor and are dependent on the stator and rotor resistances,

$R_\phi$  and  $R_\theta$  and stator and rotor RMS currents,  $i_\phi$  and  $i_\theta$ .

$$P_\Omega = i_\phi^2 R_\phi + i_\theta^2 R_\theta \quad (12)$$

There are three iron losses,  $P_{iron}$ , mechanisms, Hysteresis loss  $P_{hyst}^D$ , eddy current loss,  $P_{eddy}^D$  and anomalous loss,  $P_{anom}^D$ . Iron losses can be simplified and related to the magnetization current.

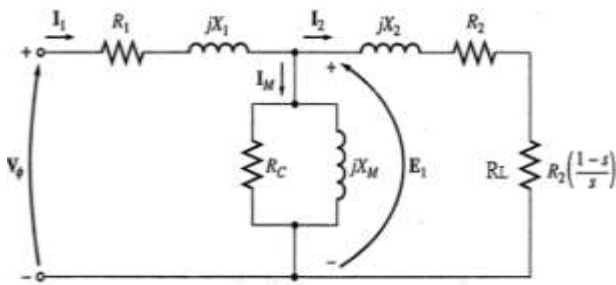


Fig.6 Induction Motor Equivalent Circuit [12]

An induction motor can be represented as a quasi-steady state equivalent circuit, Fig.6, using resistances to represent the three loss mechanisms,  $R_s$ ,  $R_r$  and  $R_c$ , representing the stator copper loss, the rotor copper loss and the iron loss respectively. The equivalent circuit can be evaluated to determine the motor losses.

### Mechanical transmission

The final stage of the drive is the mechanical transmission. The transmission consists of a gear box, usually made of up a driver gear on the motor shaft and a gear on the axle. Losses occur when one gear drives another [13]. The losses can be related to the coefficient of friction,  $\nu$ .

$$P_\sigma = P_\tau \left[ \frac{\nu}{2} (\lambda_\alpha^2 + \lambda_\beta^2) \right] \quad (13)$$

where  $P_\tau$  is the mechanical power transmitted,  $\lambda_\alpha$  is the angle of approach for the driver gear and  $\lambda_\beta$  is the angle of recess for the driver gear. Frictional losses of the mechanical transmission, including bearing losses are considered as part of the

frictional losses of the train. If frictional coefficients are determined using rundown tests, the effects of friction within the drive line are considered.

### Auxiliary loads

In general, this includes the auxiliary energy needed for the ventilation of traction motors and the traction converter cooling, but also includes the operation of the brake system for the vehicle (e.g., compressed air). As far as passenger vehicles are concerned, there is an additional energy demand to ensure passenger comfort, such as heating, lighting, and coach ventilation. This energy, which typically accounts for about 20% of the total energy consumption of a train, is supplied by the primary energy source used for traction (catenary or diesel) and delivered along the train by the auxiliary bus supply distribution [14-15].

### Traction power supply system and losses

In electrified traction systems, vehicles are powered by electricity which is supplied from a local distribution network through a traction supply system. For the purpose of analyzing the energy flows of an electrified transit system, energy flows in the electricity supply system are considered from the point of connection to the local distribution system. The traction electrical supply system includes the traction substations, the conductor system and a contact system. For DC systems, substations are located every few kilometers along the system (depending on the systems utilization and the voltage level). Substations consist of a transformer and a rectifier. Typically, 12-pulse rectifiers are used to reduce harmonic distortion on the local distribution network. Fig.7 shows a traction substation layout [16]. Transformer losses are divided into two categories, copper losses in the windings and core losses ( $P_c$ ) due to hysteresis and eddy currents losses [17]. These loss mechanisms are similar to those described for induction motors.

$$P_\chi = 3(i_1^2 R_1 + i_{2\alpha}^2 R_{2\alpha} + i_{2\beta}^2 R_{2\beta}) + P_c \quad (14)$$

where  $R_1, R_{2\alpha}$  and  $R_{2\beta}$  represent the resistance of the primary winding, and wye and delta windings of the secondary respectively.  $i_1, i_{2\alpha}$  and  $i_{2\beta}$  represent the primary and two secondary currents respectively. To simplify the equation an equivalent resistance can be used to relate the copper losses ( $P_\chi$ ) to the output current of the substation,  $i_\kappa$

$$P_\chi = i_\kappa^2 R_{e\kappa} + P_c \quad (15)$$

A 12-pulse rectifier consists of two full bridge rectifiers; each bridge contains of six diodes. At any instant four diodes are conducting, hence the losses of the rectifier can be determined by considering the losses in four diodes. Diode power losses occur due to the forward bias voltage ( $V_\xi$ ). The forward bias voltage varies with current, and can be described as

$$V_\xi = V_{\xi o}(t) + iR_\xi \quad (16)$$

These values can be derived from manufacturers data sheets, for a diode rectifier  $V_{\xi o} = 0.81$  V and  $R_\xi = 4.8$  m  $\Omega$ , [10]

$$P_\omega = 4V_\xi i_\kappa + 4R_\phi i_\kappa^2$$

(18)

where  $i_\kappa$  is the substation output current. The overall loss of the substation is determined by adding the transformer losses to the rectifier losses and can be described as a quadratic function of the substation current.

$$P_\kappa = i_\kappa^2 (R_{2e\kappa} + 4R_\phi) + 4V_\xi i_\kappa + P_c \quad (19)$$

The power loss can be related to the substation power,  $P_\kappa$

$$P_\kappa = i_\kappa^2 \left( \frac{R_{2e\kappa} + 4R_\phi}{V^2} \right) + \frac{4V_\xi}{V^2} p_{s\kappa} + P_c \quad (20)$$

This is integrated to calculate the energy dissipated.

$$P_\kappa = \int_{t_o}^t \left( i_\kappa^2 \left( \frac{R_{2e\kappa} + 4R_\phi}{V^2} \right) + \frac{4V_\xi}{V^2} p_{s\kappa} + P_c \right) dt \quad (21)$$

Power is transmitted from the traction substations to the vehicles through a conductor system, for light rail systems this is usually an overhead line catenary, some metro systems use conductor rails. Energy is dissipated

through  $I^2R$  losses in the conductors. Fig. 16 shows a double end fed section, with a single

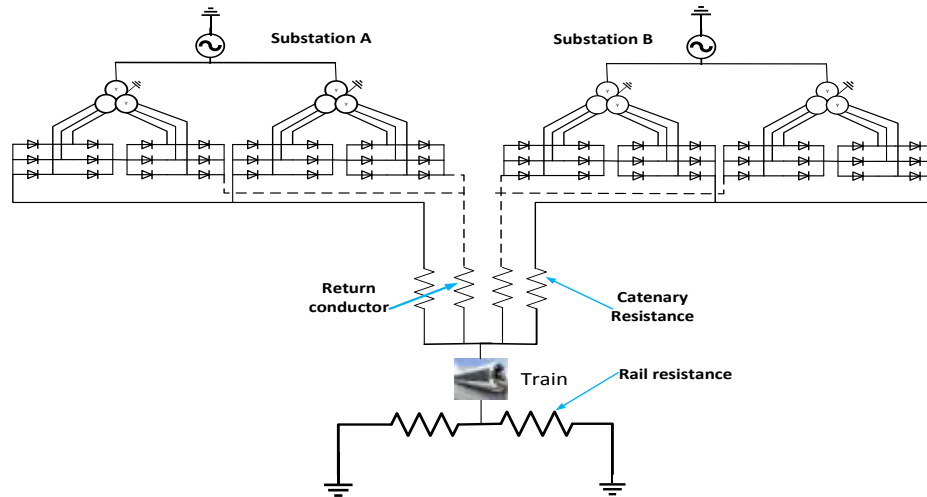


Fig.7 Addis Ababa light rail transit traction power supply network [16]

$$P_\xi(t) = V_{\xi o} i_\xi + i_\xi^2 R_\xi \quad (17)$$

The four diodes that are conducting each conduct the rated current. The power loss in the rectifier ( $P_\omega$ ) can be described as,

vehicle. Current is supplied from both substations through resistances  $R_{sA}$  and  $R_{sB}$ , and returns through  $R_{rA}$  and  $R_{rB}$ . The transmission losses in this system can be

described by adding the losses in each length of conductor.

$$P_{sL} = i_{sA}^2 R_{sA} + i_{sB}^2 R_{sB} + i_{rA}^2 R_{rA} + i_{rB}^2 R_{rB} \quad (22)$$

This can be generalized for any system and described as the sum of copper losses:

$$P_{sL} = \sum i_n^2 R_n \quad (23)$$

where  $i_n$  is the current and  $R_n$  is the resistances of sections of the supply system. To consider the energy loss, the power loss is integrated over time. The currents and resistances of each section can be described as functions of time,  $I(t)$  and  $R_n(t)$  respectively.

$$E_{sL} = \int_0^t (\sum i_n^2(t) R_n(t)) dt \quad (24)$$

### Case study: Addis Ababa light rail transit system

The analysis described in this research paper is applied to the city of Addis Ababa light rail system to determine the distribution of energy dissipation over the drive cycle shown in Fig.8 below.

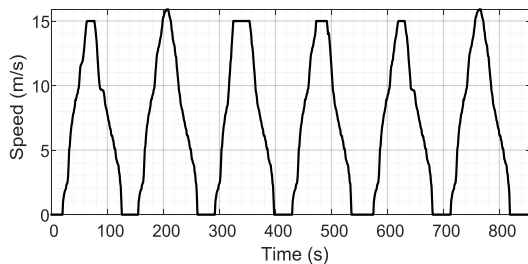


Fig.8 Speed profile of light rail transit system

frictional forces is calculated using equation (25). The case study is based on Addis Ababa light rail transit and therefore the frictional parameters of the light rail can be approximated to those of electric motor train of Addis Ababa. Davis relates the frictional forces for an electric motor train resistance,  $F_R$  to the train mass,  $m$  (kg), number of axles,  $n$ , frontal area  $A$  ( $m^2$ ) and the speed  $v$  (m/s).

$$F_R = 0.933 \sqrt{mn} + 12700 \frac{n}{m} + 8.81 \cdot 10^{-4} mv + 0.575 Av^2 \quad (25)$$

Taking parameters of the Addis Ababa light rail transit (Table 1), the coefficients of the Davis equation can be determined;  $a = 481.11$ ,  $b = 38.76$   $c = 5.75$ .

Table1. Addis Ababa light rail transit specification [18]

Mass, m	44,000 Kg
Axle, n	6
Frontal area, A	10 m <sup>2</sup>
Average speed	15 m/s <sup>2</sup>

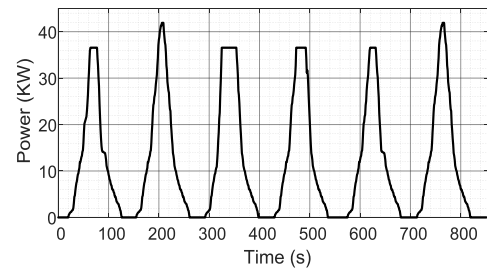


Fig.9 Frictional loss over a simple cycle

Fig.9 shows the friction power profile over drive cycle and found out that 0.551kWh/km energy is dissipated through frictional forces.

### Mechanical transmission losses

Gear box losses can be determined by using equation (13), to determine this, the mechanical power,  $P_r$  is required. The mechanical power is the power required to overcome friction and accelerate the vehicle, it is assumed that the vehicle travels on a level track, and hence no energy is required to climb a gradient.

$$P_r = P_R + P_a \quad (26)$$

Friction is calculated in the previous section. The power to accelerate the train can be calculated from equations of motion (27).

$$P_a = m.v.a \quad (27)$$

where  $m$  is the equivalent mass, which is the sum of the vehicle mass and the equivalent mass of the vehicle rotational parts. The mechanical power for the complete drive cycle is shown in Fig.10.

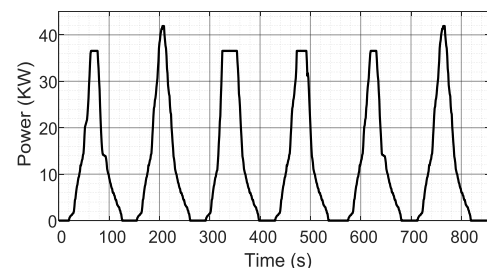


Fig.10 Mechanical power of a single train



The coefficients can be taken from [13] as  $\nu' = 0.0272$ ,  $\lambda_\alpha = 0.3691$  rads and  $\lambda_\beta = 0.3045$  rads. The total energy dissipated through the gears is 0.01kWh/ km. The energy loss profile is shown in Fig.11.

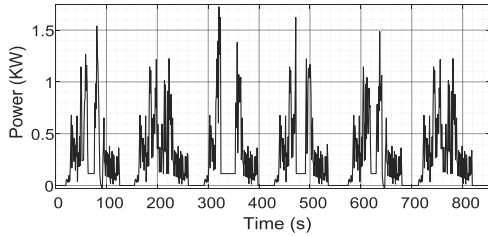


Fig.11 Power dissipated in the gears

The energy dissipated through the gears is small, and so for the purpose of analysis it is added to the frictional losses.

### Induction motor losses

The induction motor losses can be divided into three categories; ohmic losses, iron losses, frictional and wind age losses. Frictional and wind age losses are considered as frictional losses of the vehicle. An induction motor can be described as an equivalent circuit as shown in Fig.6. Where  $X_r'$  and  $R_r'$  are equivalent values. The power dissipated in  $R_r'(1-s)/s$  represents the mechanical power generated by the induction motor  $P_r$ .

$$P_r = I_1^2 R_r' \left( \frac{1-s}{s} \right) \quad (28)$$

The mechanical power is equal to the sum of the train mechanical power and the mechanical transmission loss. The power losses in the induction motor can be determined by evaluating the equivalent circuit represented in Fig.6 (for further reference a detail mathematical analysis related to induction motor losses has been performed in [19]).

The power dissipated in the resistances represents the power losses. To determine the power dissipated in the stator and rotor windings, the voltage and frequency of the supply are required. VVVF supply inverters produce a variable voltage variable frequency

supply and the voltage is proportional to the frequency.

The equivalent circuit can be solved to determine the losses in each component. The equivalent circuit cannot be solved analytically to determine the required supply voltage and frequency, and therefore the equivalent circuit should be solved numerically, using iterative steps to determine the required voltage and frequency. When the supply voltage and frequency have been determined numerically, the equivalent circuit can be solved to calculate the powers dissipated in each of the equivalent circuit resistors. For a given rotational speed, the motor will have a torque profile depending on the applied frequency and voltage. The torque at a given frequency can be calculated by evaluating the equivalent circuit. The input voltage is proportional to the frequency. The torque can be determined by finding the power dissipated through  $R_r'(1-s)/s$  of the equivalent circuit. The rotor circuit can be evaluated by substituting the supply, stator circuit and iron circuit with the venin equivalent supply, Fig.12.

$$v \rightarrow f \quad (29)$$

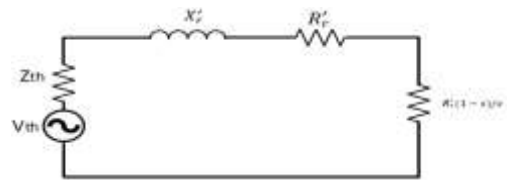


Fig.12 Rotor circuit with a Thevenin equivalent source

The parameters of the venin equivalent voltage and impedance can be calculated using equations (30) and (31).

$$V_{th} = V_m \frac{\sqrt{x_m^2 + x_c^2}}{\sqrt{R_m^2 + (x_s + x_m)^2}} \quad (30)$$

$$Z_{th} = R_{th} + jX_{th} = \frac{jx_m(R_s + jx_s)}{R_s + j(x_s + x_m)} \quad (31)$$

The torque  $T_M$  produced can be calculated using equation (32)

$$T_M = \frac{3}{\omega} \frac{V_{th} \frac{R_r}{s}}{(R_{th} + \frac{R_r}{s})^2 + (x_{th} + x_r)^2} \quad (32)$$

Fig.13 shows a motor torque profile at a given rotational speed. The applied motor frequency must lie within the stable region, and hence the minimum and maximum torques should be determined. The minimum and maximum can be determined by sweeping through the frequencies and calculating the torque each time: the point at which torque is at a minimum represents the minimum supply frequency and likewise for the maximum obtainable torque the solution must lie

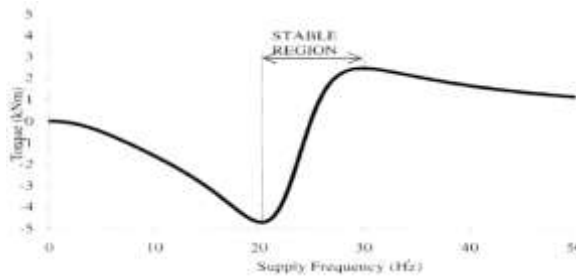


Fig.13 Motor torque profile at 78.5rad/s [20] calculated torque and the required torque, the error, decreases to an acceptable level. When the supply frequency, and therefore the supply voltage has been determined, the equivalent circuit is solved to find the power dissipated in the stator resistor, rotor resistor and core loss resistor to determine the stator copper loss, the rotor copper loss and the iron core loss respectively.

Table 2. 130 kW motor specifications [21]

Parameters	Values
Stator inductance	0.3121mH
Rotor inductance	0.3121mH
Stator resistance	0.0351Ω
Rotor resistance	0.0211Ω
Magnetizing inductance	1.2mH
Core loss resistance	215Ω

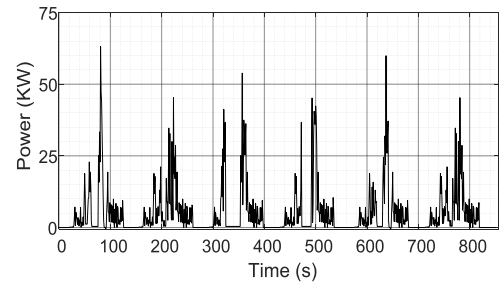


Fig.14 Induction Motor loss for the drive cycle

The total energy dissipated by each induction motor over the drive cycle is 0.083 kWh/km. This gives 0.334 kWh/km for the four induction motors. Fig.14 shows the total motor loss profile for the vehicle.

### Power Electronic Converter losses

The induction motor is driven by an inverter. An inverter consists of six IGBTs with anti-parallel diodes. Inverter losses can be divided into two categories, conduction losses (see equations (7-10)) and switching losses (see equation (11)). The conduction loss of an inverter is dependent on the motor current and whether the diode or IGBT is conducting. For the given equations (7) and (9), the values of  $V_\alpha(t)$  and  $V_\psi(t)$  are, 2.3V and 2.4 V respectively [10]. This means the conduction losses are similar, and so the analysis can be simplified by assuming all the current is conducted by the IGBTs. The conduction loss can be described as a function of the motor current,  $I_M$  as shown in equation (34) below,

$$P_{cn} = V_{\psi o} I_m + R_\psi I_m^2 \quad (34)$$

where  $V_{\psi o} = 1.83$  V and  $R_\psi = 17.73$  mΩ. The switching losses depend on the switching frequency, as shown in equation (11). The switching losses both  $\gamma_{on}$  and  $\gamma_{off}$  are functions of current and can be obtained analytically.

They are normally displayed graphically on component datasheets. The relationship can be approximated to a linear relationship as shown in equation (35),

$$P_s \approx \epsilon_\phi a I_m \quad (35)$$

where  $a$  is constant, for Siemens devices this value is calculated to be  $3 \times 10^{-4}$  V. The total loss for the inverter can be described as a function of current.

$$P_{LLS} = V_{\psi o} I_m + R_{\psi} I_m^2 + \varepsilon_{\phi} a I_m \quad (36)$$

Taking a switching frequency of 20 kHz and a DC link Voltage of 590 V, the total energy dissipated in each inverter is calculated to be 0.0402 kWh/km. The power electronics driving the motors have a total energy dissipation of 0.16 kWh/Km; Fig.15 shows the power dissipation profile for the power electronics on the vehicle.

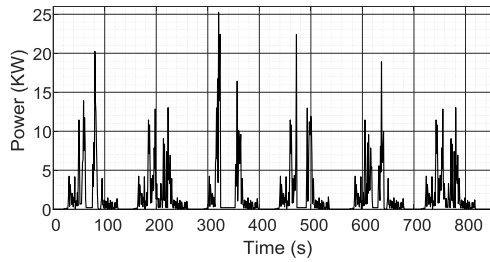


Fig.15 Vehicle Power Electronics Power loss profile

The auxiliary load for Addis Ababa light rail transit is assumed constant; measurements have indicated that the average auxiliary load is 15 kW. Over the drive cycle, 0.416 kWh/Km of energy is consumed by the auxiliary loads.

### Traction supply system losses

The traction supply system losses can be determined by analyzing the traction supply network. The network impedances are dependent on the position of the vehicle, as the lengths of conductors vary, and hence the parameters of the electrical network vary with time. The losses in the electrical network depend on the total currents, and hence are dependent on all vehicles in the system. Full analysis of system losses should consider all vehicles.

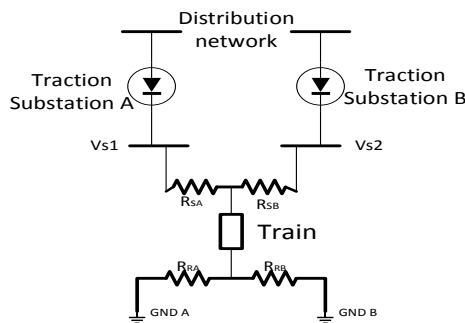


Fig.16 Electrical circuit for a double end fed section of an electrified transit system

To perform the assessment a single train a double end fed section is considered, Fig.16. The section is 1 km long. For the analysis, the train starts in the middle of the section and finishes at the end. The values of the lumped resistances are calculated from the resistance per unit length and the lengths

determined from the position of the vehicle. Values of 0.123  $\Omega$ / km and 0.0924  $\Omega$ /km are used for the resistance per unit length of the supply and return conductors respectively [22]. By applying nodal voltage analysis to the system the currents in each section of conductor can be determined. The losses can be determined and summed to find the total transmission loss as shown in equation (23). Fig.17 shows the power loss dissipation profile for the traction supply system and it is found out that 0.034 kWh/km of energy are dissipated in the supply system conductors.

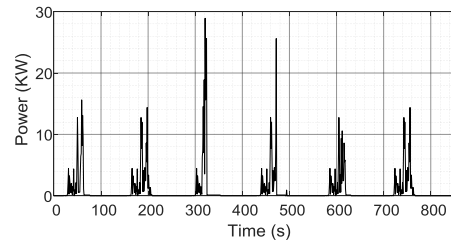


Fig.17 Traction Supply System Losses

The analysis of the traction power supply system network is also used to determine the currents drawn from each substation, which help us to calculate the substation losses. Consequently, equation (19) is used to determine this loss.

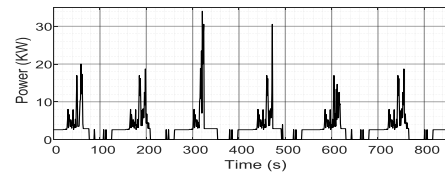


Fig.18 Traction substation losses

Resistance ( $R_{2ek}$ ) (on the DC side), is taken as 0.0233  $\Omega$  [19].  $R_{\phi}$  is the equivalent resistance of the diodes,  $V_{\xi}$  is the forward bias voltage of the diodes. These values were found to be 0.8 V and 7.7 m $\Omega$  respectively [10]. In addition to that the core losses for a 2000 kVA, is 2.6 kW. Finally, the

substation losses are approximated to be 0.236 kWh/Km as shown in Fig.18.

### Braking resistor dissipation

During braking, kinetic energy of the vehicle is transferred through the vehicle driveline; the remaining energy is dissipated through braking resistors. The total energy dissipated during the drive cycle is 1.08 kWh/km; Fig.19 shows the braking power profile.

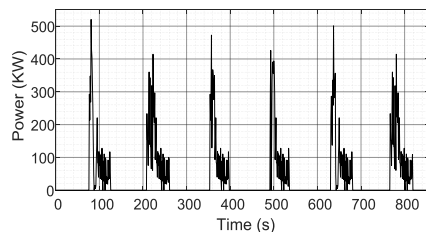


Table 3. Distribution of energy dissipation over the drive cycle.

Components	Energy loss (kWh/Km)	% of total energy loss
Frictional force	0.561	20 %
Induction motor	0.334	11.6 %
Power electronic Devices	0.161	5 %
Auxiliary	0.416	14.5 %
Traction supply losses	0.317	11 %
Braking energy	1.08	37.9 %

Table 3 shows the breakdown of energy loss over a drive cycle. The breakdown only considers the vehicle losses. Based on the analysis it is observed that 37.9 % of total energy dissipates on braking resistor and this tells us that a great deal energy could be saved if energy-storing devices were applied on Addis Ababa light rail system.

### CONCLUSIONS

Energy and environmental sustainability are becoming increasingly important as a result of expanding global urbanization. In this sense, compared to other modes of transportation such as road transport, railway plays a major role in lowering energy consumption and CO<sub>2</sub> emissions. Despite its inherent efficiency, the rail industry still consumes a significant amount of energy, making railway energy

efficiency a global priority. In this regard, Addis Ababa light rail transit systems also looking for methods to enhance their reliance on sustainable energy. Realizing the above problem, this paper has presented a detailed methodology for analyzing energy flows in a traction system of Addis Ababa light rail transit. The analysis is used to determine the energy dissipated in each component of an electrified transit system. Consequently, the study found out that 37.9% of total energy loss in the traction system is dissipated in braking. This emphasizes the potential of recovering this dissipating energy by using a certain mechanism such as implementing regenerative braking system and energy storage.

### REFERENCES

- [1]. Rohit Sharma, Peter Newman, Urban Rail and Sustainable Development Key Lessons from Hong Kong, New York, London and India for Emerging Cities, Transportation Research Procedia, Volume 26, 2017,Pages 92-105,
- [2]. Sébastien Sauvé, Sophie Bernard, Pamela Sloan, Environmental sciences, sustainable development and circular economy: Alternative concepts for trans-disciplinary research, Environmental Development, Volume 17, 2016,Pages 48-56,
- [3]. Jonas Åkerman, Anneli Kamb, Jörgen Larsson, Jonas Nässén, Low-carbon scenarios for long distance travel 2060, Transportation Research Part D: Transport and Environment, Volume 99,2021.
- [4]. Katalin Bódis, Ioannis Kougias, Arnulf J ger Waldau, Nigel Taylor, Sándor Szabó, A high resolution geospatial assessment of the roof topsolar photovoltaic potential in the European Union, Renewable and Sustainable Energy Reviews, Volume 114, 2019.
- [5]. EVENT Evaluation of Energy Efficiency Technologies for Rolling Stock and Train Operation of Railways," Institute for Futures Studies and Technology Assessment, Berlin March 2003.

- [6]. Williamson, S. Emadi, S. A. and Rajashekara, K. Comprehensive Efficiency Modeling of Electric Traction Motor Drives for Hybrid Electric Vehicle Propulsion Applications," Vehicular Technology, IEEE Transactions on, vol. 56, pp.1561-1572, 2007.
- [7]. Zhongbei Tian, Ning Zhao and Stuart Hillmansen "Traction Power Substation Load Analysis with Various Train Operating Styles and Substation Fault Modes" energies, Switzerland, June 2020.
- [8]. Kulwora wanichpong, T. Multi-train modeling and simulation integrated with traction power supply solver using simplified Newton–Raphson method. J. Mod. Transport. 23, 241–251 (2015).
- [9]. Bai Baodong and Chen Dezhi, "Inverter IGB Tloss analysis and calculation," 2013 IEEE International Conference on Industrial Technology (ICIT), Cape Town, 2013, pp. 563569.
- [10]. Siemens Semiconductor Group, "BSM 75 GB120 DN2 Datasheet" [Online]. Available:<https://static6.arrow.com/aropdfconverson/26dd3f85945dbe776875a014616b61c5e333cec1/75gb120dn2>.accessed February 2020.
- [11]. Bulent Sarlioglu, Understanding Electric Motors and Loss Mechanisms, university of Wisconsin Madison, 2016.
- [12]. Diaz, A. Saltares, R. C. Rodriguez, R. FNunez, E. I. Ortiz-Rivera and J. Gonzalez Llorente, "Induction motor equivalent circuit for dynamic simulation," 2009 IEEE International Electric Machines and Drives Conference, Miami, FL, 2009, pp. 858-863.
- [13]. Buckingham, E. Analytical Mechanics of Gears: Dover Publications, 1988.
- [14]. Erik Magni Vinberg, Energy Use in the Operational Cycle of Passenger Rail Vehicles, Master of Science Thesis Stockholm, Sweden2018
- [15]. Fisher I. and Bolton G., "Auxiliary power systems for rolling stock," in IEE Eighth Residential Course on Electric Traction Systems, 11-15 Oct. 2004, Manchester, UK,2004.
- [16]. Asegid Belay Kebede, Getachew Biru Worku,A research on regenerative braking energy recovery: A case of Addis Ababa light rail transit, transportation, Volume 8, 2021.
- [17] Mousavi, S. Shamei M., Siadatan A., Nabizadeh F. and Mirimani, S. H. "Calculation of Power Transformer Losses by Finite Element Method," 2018 IEEE Electrical Power and Energy Conference (EPEC), Toronto, ON, 2018, pp. 1-5,
- [18]. China Railway Group (CRECG), Technical Specifications of Vehicles for Addis Ababa light rail transit, internal document, 2011.
- [19]. Katsumi Yamazaki, Loss Calculation of Induction Motors Considering Harmonic Electromagnetic Field in Stator and Rotor, Electrical Engineering in Japan, Vol. 147, No.2, 2004.
- [20]. Martyn Chymera, the implementation of an energy storage system on-board a light rail traction vehicle, PhD thesis, School of Electrical and Electronic Engineering, john Rylands University, 2019.
- [21]. China Railways SS1, Ethiopian light rail transit traction motor detail specification, internal document, 2015.
- [22]. Asegid Kebede; Shimelis Atile; Demisu Legese, Harmonic Analysis of Traction Power Supply system: Case Study of Addis Ababa Light Rail Transit, IET Electrical Systems in Transportation, 25 March 2020.

# DETECTION AND RESTORATION OF CLICK DEGRADED AUDIO BASED ON HIGH-ORDER SPARSE LINEAR PREDICTION

Bisrat Derebssa<sup>1</sup>, Eneyew Adugna<sup>2</sup>, Koen Eneman<sup>3</sup> and Toon van Waterschoot<sup>4</sup>  
<sup>1,2</sup> School of Electrical and Computer Engineering, Addis Ababa Institute of Technology, Addis Ababa University, Ethiopia  
<sup>3,4</sup> Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, Belgium  
Corresponding Author's Email [bisrat@aait.edu.et](mailto:bisrat@aait.edu.et)

## ABSTRACT

*Clicks are short-duration defects that affect most archived audio media. Linear prediction (LP) modeling for the representation and restoration of audio signals that have been corrupted by click degradation has been extensively studied. The use of high-order sparse linear prediction for the restoration of click-degraded audio given the time location of samples affected by click degradation has been shown to lead to significant restoration improvement over conventional LP-based approaches. For the practical usage of such methods, the identification of the time location of samples affected by click degradation is critical. High-order sparse linear prediction has been shown to lead to better modeling of audio resulting in better restoration of click degraded archived audio. In this paper, the use of high-order sparse linear prediction for the detection and restoration of click degraded audio is proposed. Results in terms of click duration estimation, SNR improvement and perceptual audio quality show that the proposed approach based on high-order sparse linear prediction leads to better performance compared to state of the art LP-based approaches.*

**Index Terms:** Click degradation, Missing sample estimation, High-order sparse linear

*Prediction, linear prediction, Backward prediction*

## INTRODUCTION

According to [1] click degradation refers to “localized artifacts which occur at random positions in an audio signal”. These are often due to physical damages on medium and annoying to listen to. Clicks can be modeled as additive or as replacement degradation. An additive model, where the click degradation is assumed to be added to the underlying audio signal, has been shown to be acceptable for most surface defects in recording media, such as dust, dirt and small scratches [1]. A replacement model, where the degradation replaces the signal entirely for some short period of time, maybe applicable for breakages and large surface scratches which may completely destroy the underlying signal information. Generally, restoration of click-degraded audio can be seen as missing sample estimation if the underlying signal during the occurrence of the click is assumed to be lost and the time location of the click degradation is known. A method used for the restoration of click-degraded audio should only modify samples that are affected by click degradation by utilizing properties of the undegraded signal before and after the degraded signal segment. To avoid unnecessary distortion of the sample values that are not degraded a



detection stages first carried out to locate samples that are affected by click degradation. Restoration is then carried out only for the samples on these detected sample locations.

The detection of click degraded samples, in short, click detection, can be cast in a statistical framework as the detection of samples that are not generated from the same random process as the underrated audio signal [1]. From this perspective, click detection becomes equivalent to outlier detection which is a widely researched problem in the field of statistical data analysis. Some of the most widely used click detection methods are based on linear filtering and autoregressive modeling.

- **Highpass Filtering:** This approach is based on the assumption that most audio signals contain little energy at high frequencies (greater than 10 kHz), while clicks have spectral content at all frequencies. Therefore, by using a high pass filter, clicks can be enhanced relative to the underlying signal [1]. Time domain power thresholding can be used after the filtering to detect those segments of the audio signal degraded by clicks. This method is one of the earliest click detection methods used in both analog and digital audio equipment [1]. It is simple to implement with only the filter cutoff frequency and the detection threshold as parameters. The method will fail if the clicks are band-limited or if the signal has high frequency content, such as high-pitched musical instruments.
- **Autoregressive (AR) model-based click detection:** Model-based click detection methods use prior information about the underrated signal and the clicks into the detection procedure in the form of hypothesized signal models. In this approach, the underrated audio signal is assumed to be drawn from a short-term

stationary process while the clicks are assumed to behave as impulsive noise. This AR modeling is very effective for human speech representation and is the basis for different audio signal representation schemes ranging from audio encoding, audio compression and audio feature extraction [2].

For AR modeling of an underrated audio signal, the prediction error is expected to take on small values while the prediction error will be large if an impulsive noise that is not correlated with the underrated audio signal replaces the signal. Therefore, clicks can be detected by inverse filtering an audio signal using an AR model prediction error filter (PEF) and by thresholding the prediction error [1], [3], [4], [5], [6]. The limitations of this approach and researches conducted to address these are discussed below.

The PEF will spread a single impulse over future samples thereby creating interference with other impulses located in close proximity. This may make detection threshold selection problematic.

It is difficult to estimate the end time of a click due to the forward smearing effect of the PEF. Backward prediction has been used successfully to resolve this problem [1].

If the underlying audio signal is not produced by an AR process, the AR model may not well represent the signal and the prediction error may be large. In this case, false positives may be reported. This may be the case for voiced speech and high-pitched musical notes where the AR model order may not be large enough. Autoregressive moving average (ARMA) modeling and high-order linear prediction have been proposed to better represent musical signals [1], [2], [7].

Several methods have been proposed for the restoration of click-degraded audio. The Least Squares (LS) estimation of the AR model coefficients, in the sequel referred to as linear prediction (LP) minimizes the square error (MSE) criterion assuming that the AR model excitation signal has a Gaussian distribution. It assumes that the undegraded audio signal is generated by passing a white noise excitation through an all-pole filter and that the click-degraded samples are mutually independent and drawn from a Gaussian zero-mean process. The click-degraded samples can then be restored by LP-based interpolation from a priori knowledge of the LP coefficients of the undegraded audio signal, of the undegraded samples and of the time location of the click-degraded samples.

One of the limitations of the LP-based interpolator is the unavailability of the LP coefficients of the undegraded signal. An iterative procedure for estimating the LP coefficients and then interpolating the missing samples was proposed by Janssen et al. [8] applying the Levinson-Durbin recursion in each iteration. Even though this approach works well for unvoiced speech [7], it is not suitable for music and voiced speech, where the AR model excitation is quasi-periodic and spiky [8]. For voiced speech and music, the minimization of the MSE, i.e., the  $l_2$ -norm of the LP vector residual puts more emphasis on the periodic spikes of the residual [2]. This problem could be resolved by including a pitch predictor in the AR model to estimate long-term correlation. However, this ignores the interaction between the long-term and short-term predictors, leading to a sub-optimal result. Joint optimization of the long-term and short-term predictors was proposed in [9]. Recently a method for the joint detection and restoration of click-degraded archived audio that uses a joint evaluation of signal prediction errors and leave-one-out signal interpolation errors was proposed [6]. It is based on thresholding the prediction error for click detection followed by multi-step ahead signal prediction. In this approach, the LP

coefficients are estimated by the Levinson-Durbin recursion and restorations done by LS interpolation. The use of the conventional LP, i.e., short-term LP may limit the performance of this approach.

A better decoupling between the LP-based modeling of spectral envelope and pitch harmonics has been reported by using high-order sparse linear prediction (HOSpLP) [7],[10], [11]. In our previous work [12], [13] the use of  $l_1$ -norm regularized and  $l_0$ -norm regularized HOSpLP for the restoration of click-degraded audio given the time location of the click degradations has been investigated. Extensive simulations showed that the use of HOSpLP results in improved restoration performance compared to [8] in terms of signal-to-noise ratio (SNR) and perceptual evaluation of audio quality (PEAQ). In this paper the use of HOSpLP coefficients for the detection of click-degraded samples and restoration of these samples that works for both speech and music without priori on the type of audio is proposed. This will significantly decrease the need for manual annotation (speech vs. music) and segmentation (undegraded vs. degraded segments) needed for practical application.

The contribution of this paper is twofold. First, we extend the use of HOSpLP, proposed in [12], [13] for the restoration of click-degraded audio, to click detection. Second, a unified detection and restoration method based on HOSpLP coefficients is proposed. Simulation results are included to show the superior performance of the proposed HOSpLP coefficients for detection as well as restoration of click-degraded audio in comparison to state-of-the-art LP-based approach. The organization of the paper is as follows. Section informally discusses the HOSpLP coefficients considering both  $l_1$ -norm and  $l_0$ -norm regularization to induce sparsity. Section III

discusses the problem of click detection and proposes two click detection approaches based on HOSpLPcoefficients. Section IV unifies the detection and restoration problem. Section V discusses the data, the artificial click degradation and the performance measures used in the simulations. Section VI presents simulation results on click detection and restoration and a comparative performance evaluation to state-of-the-art approaches. Finally, Section VII concludes the paper.

### HIGH-ORDER SPARSE LINEAR PREDICTION

Linear prediction (LP) is a well-understood and widely used method for the analysis, modeling, and coding of speech signals [2]. Its success is due to its alignment with the source filter model of the speech generation process [14]. It has been shown that a slowly time-varying, low-order all-pole filter can be used to model the vocal tract. The glottal excitation is modeled as either an impulse train for voiced sounds or a white noise sequence for unvoiced sounds. The purpose of all-pole modeling through LP is to obtain a short-term predictor that characterizes the spectral envelope of the vocal tract response.

The LP coefficient vector  $\mathbf{a}$  can be estimated for a frame of observed samples  $\mathbf{x}$  by solving the following optimization problem [7]

$$\hat{\mathbf{a}} = \operatorname{argmin}_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma \|\mathbf{a}\|_k^k \quad (1)$$

Where

$\mathbf{X}$	=	$\begin{bmatrix} x_{N_1-1} & \dots & x_{N_1-P} \\ \vdots & \ddots & \vdots \\ x_{N_2-1} & \dots & x_{N_2-P} \end{bmatrix}$
$\mathbf{a}$	=	$[a_1 \dots a_P]$
$\mathbf{x}$	=	$[x_{N_1} \dots x_{N_2}]$
P	is	the order of the LP model
$N_1$	are	the start and end indices of the

and $N_2$		frame under consideration.
$\gamma$	is	a regularization parameter

The  $l_p$ -norm operator  $\|\cdot\|_p$  is defined as

$$= \left( \sum_{n=N_1}^{N_2} |x_n|^p \right)^{\frac{1}{p}} \quad (2)$$

For conventional LP solved via the Levinson-Durbin algorithm, the  $l_2$ -norm is used,  $p = 2$ , and no structure on the coefficient vector is imposed,  $\gamma = 0$ . Furthermore, the prediction order is usually set to a small value corresponding to twice of the number of formant frequencies to be modeled. Even though such modeling works well for unvoiced speech where the excitation can be modeled as white noise [7], it is not a good model for music and voiced speech, where the excitation is quasi-periodic and spiky [8]. For voiced speech, the excitation is appropriately modeled as periodic pulse train corresponding to the glottal output. As such the minimization of the  $l_2$ -norm of the residual puts more emphasis on the periodic spikes of the residual [2]. As a result, it tradeoff short-term prediction, i.e., spectral envelope, estimation accuracy against the long-term prediction, i.e., pitch estimation accuracy [2]. As the aim of conventional LP is to model the vocal tract and not the glottal excitation, this leads to a suboptimal solution.

For musical sounds or tonal audio for which the signal contains a finite number of dominant frequency components, the LP model is much less popular than in speech analysis as the generation of musical sounds is dependent on the instruments used [2]. This makes it hard to use a generic audio signal generation model [2]. In addition, each polyphonic audio signal should be modeled using multiple source-filter models [2], [14]. In the absence of noise, by using a

model order which is twice the number of tonal components, LP can be used to estimate the spectral peaks. In practice, noise is always present that may be due to imperfections in the tonal behavior, or lack of tonal behavior, finite precision arithmetic, finite-length data windowing or background or sensor noise. Therefore, such LP signal estimates are very often poor. In [2] extensive simulations were conducted to assess the performance of conventional and alternative LP models for tonal audio analysis in the presence of noise. It was reported that high-order all-pole models are better suited to the audio LP problem albeit being impractically complex in many applications due to the excessive number of LP coefficients.

One of the most recent approaches to LP is sparse linear prediction (SpLP), which takes into consideration the sparsity of the residual and the LP coefficients. When applied to high-order all-pole models, SpLP is referred to as high-order sparse linear prediction (HOSpLP). A better decoupling between the spectral envelope and pitch estimation has been reported by using HOSpLP [7], [10], [11], [12], [13]. While the high-order all-pole method used in [2] minimize the  $l_2$ -norm of the residual to obtain the LP coefficients, the HOSpLP methods impose sparsity of the residual and the coefficient vector in the optimization problem.

#### A. $l_1$ -norm regularized HOSpLP

By imposing sparsity of the residual in the LP problem formulation the emphasis on outliers in the solution to (1) can be decreased [7]. That is by considering a sparsity-inducing norm of the residual vector instead of the  $l_2$ -norm. The convex relaxation of the ' $l_0$ -norm' cardinality problem has been proposed to lead to a sparser residual [7]

In addition, by using a high-order all-pole model and imposing sparsely of the coefficient vector in (1), by setting  $\gamma=0$  and  $k = 1$ , joint estimation of the short-term predictor and the long-term predictor can be achieved [7] as in (3).

$$\hat{\mathbf{a}} = \operatorname{argmin}_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 \quad (3)$$

This results from the observation that a cascade of a long-term and short-term predictor results in a filter that has few non-zero coefficients [14]. Therefore, the sparsity of the coefficient vector can be used to regularize the solution. The purpose of the HOSpLP coefficients obtained by solving (4) is to model the whole spectrum, i.e., the pitch related harmonics and the spectral envelope.

$$= \operatorname{argmin}_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma \|\mathbf{a}\|_1 \quad (4)$$

The problem in (4) is convex but not differentiable. However, it can be solved via splitting methods such as the alternating direction method of multipliers (ADMM) by reformulating the problem as a basis pursuit problem [15]. The regularization parameter,  $\gamma$ , determines the trade-off between the sparsity of the predictor coefficients and the sparsity of the residual. The modified L-curve [16] has been used to obtain an optimum value for the regularization parameter in [11]. In [11] an adaptive algorithm was proposed for estimating the regularization parameter based on the observation that the optimal  $\gamma$  is related to the pitch gain.

To solve the problem of obtaining the short-term and long-term predictors from a HOSpLP coefficient vector,  $\mathbf{a}$ , the first few,  $N_f$ , coefficients of the HOSpLP coefficient vector been used to represent the short-term predictor in [7]. After this, a polynomial factorization can be carried out to obtain the long-term predictor after selection of the number of taps in the long-term predictor, typically  $N_p=1$  or  $N_p=3$ .

The use of the  $l_1$ -norm in HOSpLP has been shown to outperform conventional LP in the estimation of spectral envelope, sparse LP coefficients and sparse residual [7]. With regards to stability of the obtained short-term filters, it has been shown in [7] that the percentage of unstable filters is very low (around 2%) with “mild” instability.

### *B. $l_0$ -norm regularized HOSpLP*

The prior knowledge of the structure of the coefficient vector resulting from cascading a long-term and short-term predictors can also be incorporated in the HOSpLP optimization problem as (5) [13],

$$\hat{\mathbf{a}} = \operatorname{argmin}_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 \text{ s.t. } \|\mathbf{a}\|_0 \leq \Psi \quad (5)$$

Where  $\Psi$  is the sum of the filter order of the long-term and short-term predictors.

This formulation does not impose a restrictive structure on the coefficient vector except that the coefficient vector has a fixed maximum number of non-zero coefficients. As such, it can give emphasis to the formant filter coefficients if the signaling the frame is composed of unvoiced speech and to the pitch or tonal components if the frame is composed of voiced speech or music. In [13] it was shown that the coefficients obtained by solving (5) correspond to the short-term and long-term predictor. As the location of the non-zero coefficients is neither incorporated into (5) nor dependent on a pitch predictor, prior information regarding the type of signal is not needed. In addition, the structure of the coefficient vector can change from frame to frame if the signal is composed of both speech and music.

It should be mentioned that the use of the  $l_1$ -norm of the residual in (5) is expected to lead to better results as compared to  $l_2$ -norm. However,  $l_1$ -norm of the residual in

(5) is difficult to solve efficiently. Problem (5) is non-convex [17] which means that it may have several local minima and its convex relaxation, the least absolute shrinkage and selection operation (LASSO), obtained with  $p=2$  and  $k=1$  in (1), is typically solved instead [17],[18]. Nevertheless, proximal gradient methods can efficiently solve (5) if a good initialization is given, e.g., the solution of LASSO [18]. In recent work, Antonello et. al [18] developed the Structured Optimization package for the Julia programming language that can solve (5) in a reasonable time. This package is used in this work to obtain  $l_0$ -norm regularized HOSpLP coefficient vector.

## **CLICK DETECTION**

In practice the time location of the click degradation is not known a priori, therefore click detection methods are needed. One of the most widely used click detection approaches consists in energy thresholding of the LP residual [1]. This approach is based on the assumption that the click degradations not generated from the same AR random process as the undegraded audio signal. Therefore, in the presence of click degradation the energy of the LP residual in that time frame will be much larger than the energy of the residual when click degradation is not present. It has been shown in other applications that significant improvement in noise detectability can be achieved by transforming the noisy speech to the excitation domain of the speech signal [19].

In LP-based click detection methods, the energy of the LP residual at each sample is compared with an average residual energy of the frame as follows,

For  $n = 1$  to  $N$

- 1) Calculate LP residual:  $\epsilon_n = x_n - \sum_{j=1}^P \hat{a}_j x_{n-j}$
- 2) if  $|\epsilon_n| \geq K\sigma_e$ , then  $\mathbf{i}_n = 1$ , else  $\mathbf{i}_n = 0$

Where

$\sigma_e^2$	is the variance of the LP residual,
$K$	is a detection threshold,
$N$	is the frame length,
$\mathbf{i}$	is a vector representing the presence or absence of click degradation at each sample value, $\mathbf{i}_n = 1$ represents presence and $\mathbf{i}_n = 0$ represents absence of click degradation at the $n^{th}$ sample.

In this approach the start of click degradation is accurately estimated [1]. However, the end of a click degradation cannot be accurately estimated due to the forward smearing effect over  $P + 1$  samples, where  $P$  is the order of the AR model. To detect the end of a click, a moving average filter can be applied to see when the residual variance in a local window has energy lower than the threshold (or some scaled version of the threshold). However, this requires a precise tuning of the threshold and local window size to detect the end of a click degradation.

When impulses are present in close vicinity to each other their impulse responses resulting from filtering with the PEF may add constructively to give a false detection or cancel one another out [1]. In general, threshold selection is difficult when impulses of differing amplitudes are present. The use of the backward prediction error for the detection of clicks has been proposed in [1], [20].

This method takes advantage of the accurate LP-based start click identification. In this approach, once a click is detected and its start location identified, the backward prediction error is then used to detect the end of the click. By assuming that the time-reversed

signal can be reasonably modeled as an AR process, the energy of the LP residual of the time-reversed signal near the identified click start location is evaluated to detect the end of the click degradation. The backward prediction error is defined as

$$\epsilon_n^b = x_n - \sum_{i=1}^P b_i x_{n+i} \quad (6)$$

When these coefficients are obtained by using the conventional LP, the backward prediction error is composed of spikes due to the quasi-periodic excitation for voiced speech and music. This makes it difficult to select a threshold for the detection of the end of clicks without incorrectly selecting spikes due to the quasi-periodic excitation.

In this paper, the HOSpLP coefficients are used in click detection, see Algorithm 1, by exploiting the fact that the short-term and the long-term predictors can be jointly estimated using HOSpLP leading to a residual that has less spiky nature due to the quasi-periodic excitation [7].

As such, the backward prediction error in a local window near the identified click start can be used to estimate the end of the click without significantly being affected by a spiky residual. To avoid mislabeling undegraded samples between two click degradations that are close together, the backward prediction error is checked to be greater than the threshold in local window around the detected click start.

**Algorithm 1** Backward prediction using HOSpLP model

```

1: procedure BACKWARD_PRED_HOSPLP
2:   Input:  $\mathbf{x}, P, \gamma, R, K, N$ 
3:   Output:  $\mathbf{c}$ 
4:    $\hat{\mathbf{a}} = \text{COEFFICIENT}(\mathbf{x}, P, R, \gamma, \zeta)$ ;
5:    $\epsilon^x = \text{RESIDUE}(\mathbf{x}, \hat{\mathbf{a}})$ ;
6:    $\sigma_e = \text{STANDARD\_DEVIATION}(\epsilon^x)$ ;
7:   for  $n = 1$  to  $N$  do
8:     if  $(|\epsilon_n| \leq K\sigma_e)$ , break;
9:     else  $c_n = 1$ ;
10:     $\hat{\mathbf{b}} = \text{COEFFICIENTS}(\mathbf{x}^B, P, R, \gamma, \zeta)$ ;
11:     $\epsilon^b = \text{RESIDUE}(\mathbf{x}^B, \hat{\mathbf{b}})$ ;
12:    for  $l = n$  to  $n + k_{max}$  do
13:      if  $(|\epsilon_l^b| \geq K\sigma_e)$   $c_l = 1$ ; continue;
14:      for  $j = l$  to  $l + W$  do
15:        if  $(|\epsilon_j^b| \geq K\sigma_e)$   $c_{l,j} = 1$ ;  $l = j$ ; break;
16:      end
17:    if  $(j \geq l + W)$   $n = l$ ; continue;
18:  end
19:  Return

```

Where,

$\mathbf{x}$	is the click degraded signal vector;
$\mathbf{x}_B$	is the time-reversed click degraded signal vector;
$\mathbf{I}$	is the estimated location of click;
$K$	is the threshold value;
$N$	is the number of samples in each frame;
$R$	is the maximum number of ADMM iterations for $l1$ -norm HOSpLP;
$W$	is a local window size;
$\gamma$	is the regularization parameter for $l1$ -norm HOSpLP;
$\zeta$	is the residual stopping criterion for ADMM algorithm in $l1$ -norm HOSpLP.

The function  $\text{COEFFICIENTS}(\mathbf{x}, P, R, \gamma)$  obtains the LP coefficients as follows. The function  $\text{RESIDUE}(\mathbf{x}, \hat{\mathbf{a}})$  obtains the residual error by inverse filtering the signal with a AR filter with coefficients  $\hat{\mathbf{a}}$ .

- **$l1$ -norm regularized HOSpLP:** the ADMM algorithm for solving the  $l1$ -norm regularized problem [15] is used to obtain the HOSpLP coefficients [12].

- **$l0$ -norm regularized HOSpLP:** the  $l0$ -norm regularized problem (5) is solved via the Structured Optimization Julia package to obtain the HOSpLP coefficients [18].

#### IV. UNIFIED APPROACH FOR DETECTION AND RESTORATION OF CLICK-DEGRADED AUDIO

In this section, a unified approach is proposed that detects the location of click degraded-samples and restores these samples by using the HOSpLP coefficients without a prior knowledge on the type of audio and the time location and duration of the click degradation.

##### A. Detection and restoration by using backward prediction and Janssen iteration

Initially, the backward prediction based on  $l0$ -norm regularized HOSpLP coefficients is used to detect samples degraded by click degradation. Then these samples are restored by an iterative algorithm, see Algorithm 2, similar to the Janssen iteration [8], [17] but using  $l0$ -norm regularized HOSpLP for the restoration as this is shown to provide the best signal restoration performance [13].

**Algorithm 2** Detection and Restoration using backward prediction and Janssen restoration based on HOSpLP model

```

1: procedure RESTORATION_HOSPLP
2:   Input:  $\mathbf{x}, P, \gamma, R, K, N, L, Q, \zeta$ 
3:   Output:  $\mathbf{y}$ 
4:    $\hat{\mathbf{c}} = \text{BACKWARD\_PRED\_HOSPLP}(\mathbf{x}, P, \gamma, R, Q, N)$ ;
5:    $h = 1$ ;  $g = 1$ 
6:   for  $i = 1$  to  $N$  do
7:     if  $(\hat{c}_i == 1)$   $\mathbf{v}_h = \mathbf{i}$ ;  $h = h + 1$ ;
8:     else  $\mathbf{u}_g = \mathbf{i}$ ;  $g = g + 1$ ;
9:   end
10:   $\Theta = [\mathbf{v}_1 \mathbf{1}_{1 \times N} - \mathbf{1}_{M \times 1} [1, 2, \dots, N]]$ ;
11:   $\hat{\mathbf{x}}_u = \mathbf{x}_u$ ;  $\hat{\mathbf{x}}_v = 0$ ;  $\Phi = \mathbf{0}_{M \times N}$ ;  $l = 0$ ;
12:  for  $l \leq L$  do
13:     $\hat{\mathbf{a}} = \text{COEFFICIENT}(\hat{\mathbf{x}}, P, \gamma, R, \zeta)$ ;
14:     $\mathbf{b} = [1 \quad -\hat{\mathbf{a}}^T] \mathbf{A}$ ;
15:     $\Phi_{i,j} = \mathbf{b}_{\Theta_{i,j}+1}$ ;  $\forall i, j : \Theta_{i,j} \leq P$ 
16:     $\hat{\mathbf{x}} = -\Phi_{(1:M,e)}^{-1} \Phi_{(1:M,v)} \mathbf{b}_v$ ;
17:     $l = l + 1$ ;
18:  end
19:  Return

```

Where

$$\mathbf{A} = \begin{bmatrix} 1 & -\hat{a}_1 & -\hat{a}_2 & \cdots & -\hat{a}_P \\ -\hat{a}_1 & -\hat{a}_2 & \cdots & -\hat{a}_P & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\hat{a}_P & 0 & 0 & \cdots & 0 \end{bmatrix};$$

$L$  is the number of Janssen iterations.



### B. Benchmark method incorporating HOSpLP coefficients

As a comparison, a recently proposed method by Ciołek et al. [6] for the joint detection and restoration of click-degraded archived audio that uses a joint evaluation of signal prediction errors and leave-one-out signal interpolation errors is used. It is based on thresholding the forward prediction error for click detection followed by multi-step-ahead prediction for restoration.

A click start is detected when the absolute prediction error is larger than and a click end is detected if the residual at  $k_0$  iteration is smaller than a threshold and consecutive residuals are smaller than same threshold. In this approach, the LP coefficients are estimated by the Levinson-Durbin recursion and restoration is done by LS interpolation [21]. The use of the conventional LP may limit the performance of this approach due to the limited capability to model pitch and tonal components. We propose to use HOSpLP coefficients in this method by using the  $l_1$ -norm regularized HOSpLP coefficients instead of using the conventional LP coefficients solved via the Levinson-Durbin recursion. Algorithm 3 shows a simplified algorithm to illustrate where the HOSpLP coefficients to be used. The code for the original implementation is available in [22]. The reason the  $l_1$ -norm regularized HOSpLP is used instead of  $l_0$ -norm regularized HOSpLP is due to the fact that the  $l_0$ -norm regularized HOSpLP coefficients are solved by using the Structured Optimization package of Julia programming language, whereas the original code for Ciołek's method is in MATLAB. It should be mentioned that the use of HOSpLP coefficients in this method leads to significant computational cost as it yields to a solution to an iterative problem nested in another iterative problem, i.e., re-estimating the HOSpLP coefficients. The restoration is

done by using the LS interpolation method as used in their original work.

The function  $\text{COEFFICIENTS}(\hat{\mathbf{x}}, P, M, \gamma, \zeta)$  obtains the LP coefficients using Levinson-Durbin in the original method [6] and using ADMM in our proposed  $l_1$ -norm regularized HOSpLP variation of [6].

**Algorithm 3** Iterative detection and restoration via leave-one-out interpolation [6] by incorporating HOSpLP model

---

```

1: procedure CIOLEK_HOSpLP
2:   Input:  $\mathbf{x}, P, M, \gamma, K, N$ 
3:   Output:  $\mathbf{y}, I$ 
4:    $\hat{\mathbf{x}} = \mathbf{x}$ ;
5:    $\hat{\mathbf{a}} = \text{COEFFICIENT}(\mathbf{x}, P, R, \gamma, \zeta)$ ;
6:    $\epsilon^x = \text{RESIDUE}(\mathbf{x}, \hat{\mathbf{a}})$ ;
7:    $\sigma_e = \text{STANDARD\_DEVIATION}(\epsilon^x)$ ;
8:   for  $n = 1$  to  $N$  do
9:     if  $(|\epsilon_n| \leq K\sigma_e)$  continue;
10:     $i_n = 1$ ;
11:     $\hat{\mathbf{x}} = \text{Leave\_One\_Out\_Interpolation}(\hat{\mathbf{x}}, n, \hat{\mathbf{a}})$ ;
12:     $\hat{\mathbf{a}} = \text{COEFFICIENTS}(\hat{\mathbf{x}}, P, M, \gamma, \zeta)$ ;
13:     $\epsilon_n = \hat{x}_n - \sum_{j=1}^P \hat{a}_j \hat{x}_{n-j}$ ;
14:    if  $\exists l \in \{0, \dots, k_0\} : |\epsilon_{n-l}| \geq K\sigma_e$  continue;
15:  end
16:  Return
    
```

---

## SIMULATION SETUP

### A. Data used

To fairly assess the detection and restoration performance of the proposed methods, the experiments were conducted using speech (male and female) and music (singing voice and instrumental) from the Archimedes dataset [23]. In order to have comparable degradations among all signals, each signal is normalized so that the maximum amplitude is 1. Five male and five female speech from different speakers are taken. For each speech simulation is done on 100 frames each 32.5 ms. The result is then averaged among these. Similarly, for music 2 male singing voices, 2 female singing voice, 2 instrumental audio and 4 audio consisting of singing voice and instrument are used.

### B. Click Degradation Model

Usually, the start, duration and amplitude of each click degradation is modeled probabilistically. Different probability distributions for the time between impulses and for their amplitudes can be used [1], [24]. In this

work, the time location of click degradation was assumed to be uniformly distributed as the causes of click degradation are not correlated with the audio signal. As such, click degradations can occur at any location irrespective of previous click degradation location and the samples during the occurrence of click were replaced with zero-mean Gaussian noise to obtain a click degraded signal. The standard deviation of the click degradation is set as twice the standard deviation of the audio signal. The impact of the click degradation variance on the detection and restoration performance of the various methods is investigated in Section VI.

### C. Performance Measures

To evaluate click detection accuracy, the normalized MSE in click duration estimation for each data set and for a given click duration as shown in (7) is used.

$$NMSE = \sum_{h=1}^H \frac{|T_{click}(h) - \hat{T}_{click}(h)|^2}{|T_{click}(h)|^2} \quad (7)$$

where

$T_{click}$	is the actual click duration;
$\hat{T}_{click}$	is the estimated click duration;
$H$	is the total number of audio files for each dataset.

To evaluate the restoration performance, the Signal-to-noise ratio (SNR) and perceptual evaluation of audio quality (PEAQ) are used. The SNR is evaluated over the entire duration of the signal to also take into account unnecessary interpolation that may result from incorrect click detection.

$$SNR = \sum_{h=1}^H 10 \log_{10} \frac{|x(h)|^2}{|x(h) - \hat{x}(h)|^2} \quad (8)$$

Where  $x$  is a vector of the undegraded audio and  $\hat{x}$  is a vector of the restored audio.

PEAQ is used to assess the subjective quality of the restored audio signal [25]. It predicts the basic audio quality of a signal with respect to a reference signal by modeling the psychoacoustic properties of the human auditory system. It has a range of 0 to -4: 0 representing imperceptible

distortion while -4 means very annoying distortion. PEAQ has been used for the assessment of click-degraded audio restoration in [5] and [6]. The PEAQ implementation in [26] is used in this research.

## RESULTS AND DISCUSSION

The backward prediction and iterative forward prediction methods are based on thresholding the absolute value residual, backward prediction error and forward prediction error respectively, where the threshold values is not signal dependent and does not require rigorous tuning. In both cases, different threshold values were tested and a value of  $K = 3$  led to the best results in agreement with the “3-sigma” rule [6]. The parameters used during the simulations are shown in Table I.

Table I: Simulation Parameters

No	Description	Value
1	Sampling frequency	44.1kHz and 8kHz
2	Frame size	32.5 ms
3	Conventional LP order	12
4	HOSpLP order	half of frame size
5	Number of non-zero $l0$ -norm regularized HOSpLP coefficients	20
6	Artificial click duration	0.4536ms - 2.268ms
7	Local window size, $k_{max}$	5

### A. Click Detection Performance

1) *Estimation of start of click*: The backward prediction based click detection is heavily dependent on correct estimation of the start of the click degradation. To evaluate the performance of the backward prediction based click detection in the estimation of the start of the click, average absolute error in estimating the click start is shown in Figure 1 by using conventional LP and HOSpLP coefficients in the backward prediction method. The method proposed by Ciolek et.al. is also taken as a benchmark.

It is seen that Ciolek's method leads to the best estimation of the start of the click. However, note that at 44.1 kHz sampling frequency, 0.0227 ms is 1 samples, as such the backward prediction method on average leads to click start error of 1 samples only. The conventional LP and HOSpLP coefficients perform similarly in the estimation of the start of the click. The absolute error of estimation is on average 0.0227 ms, i.e. 1 sample at 44.1 kHz sampling frequency, for click degradation of duration up to 2.268 ms or 100 samples.

2) *Estimation of click duration:* Figure 2 shows the NMSE for click duration estimation for speech and music by using backward prediction based on conventional LP and HOSpLP coefficients and by using Ciolek's method. It is observed that for click duration less than 1 ms, the backward prediction based click detection fails entirely. However, for longer click durations the backward prediction based on HOSpLP leads to superior click duration estimation performance for music.

This is in agreement with the modeling assumption made regarding the HOSpLP coefficients for music. It is noted that for music at 44.1 kHz sampling frequency, even though Ciolek's method leads to superior identification of the click start, its estimation of the click duration is inferior to the backward prediction based method for both conventional LP and HOSpLP coefficients.

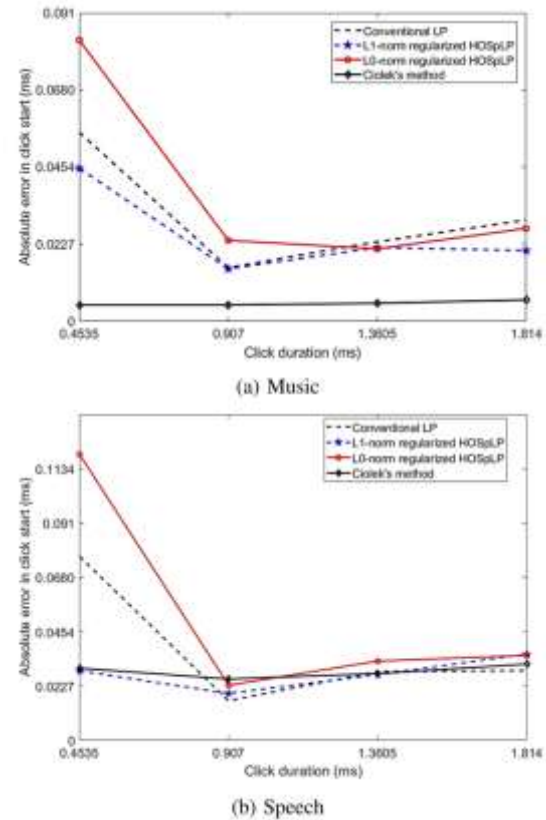


Figure 1: Absolute error in click start estimation using backward prediction using HOSpLP coefficients.

To see the impact of the sampling frequency on the click estimation of the methods, similar experiments were conducted for audio sampled at 8 kHz. Figure 3 shows the NMSE for click duration estimation for speech and music by using backward prediction based on conventional LP and HOSpLP and by using Ciolek's method for a wide range of click durations.

It is observed that for long click durations (longer than 4 ms), all methods yield similar detection performance. However, as the click duration decreases, the conventional LP and  $l_1$ -norm regularized HOSpLP accuracy decreases significantly.

The use of backward prediction with  $l_0$ -norm regularized HOSpLP coefficients leads to the best click duration estimation results for all click durations, except for very short click durations (less than 1 ms) where all methods fail. For music, it is seen that the  $l_1$ -norm regularized HOSpLP performs best for long click durations. This performance of the backward prediction

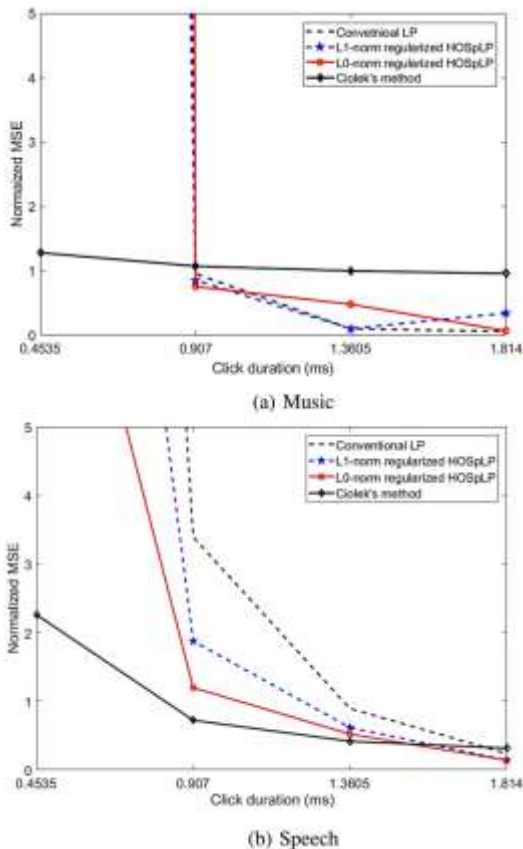


Figure 2: Performance of click duration estimation at 44.1 kHz sampling frequency.

method with HOSpLP is consistent at both sampling frequencies where as Ciolek's method leads to inferior performance as the sampling frequency is increased.

## B. Detection and Restoration performance

To measure the unified detection and restoration performance, the artificially click degraded audio was restored by using the proposed Algorithm 2 and state-of-the-art Algorithm 3 then the SNR was computed and averaged for each dataset. No information regarding the location and duration of the click degradation is used in any of the methods. Figure 4 shows the results of the detection and restoration for audio sampled at 44.1 kHz.

For audio sampled at frequency of 44.1 kHz the backward prediction method with HOSpLP coefficients leads to superior restoration performance as compared to Ciolek's method. This is attributed to the superior click duration estimation performance of the proposed backward prediction method with HOSpLP coefficients as compared to Ciolek's method.

It is also noted that the use of HOSpLP coefficients in Ciolek's method leads to improvement in restoration performance as compared to conventional LP in Ciolek's method. The improvement in SNR by the HOSpLP based methods is observed to be higher in music as compared to speech.

This can also be attributed to the superior modeling capability of HOSpLP coefficients in the case of music. This has been also seen to be the case in HOSpLP coefficient based restoration methods as reported in our previous works [13] and [12].

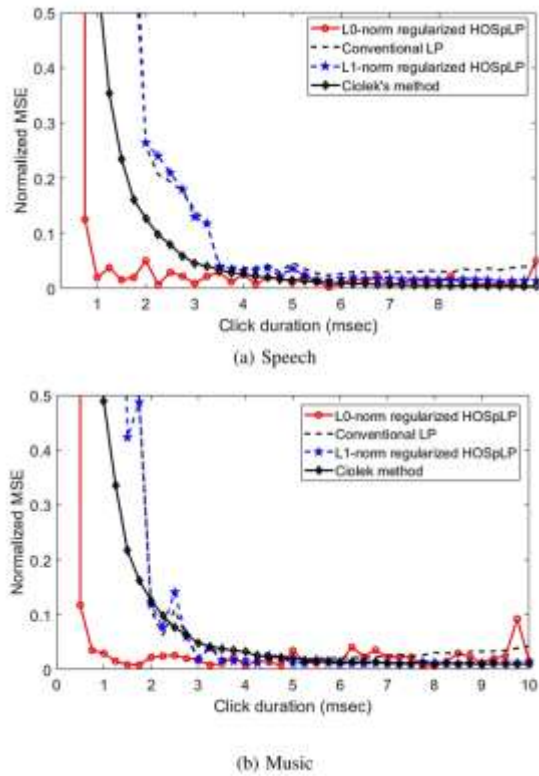


Figure 3: Performance of click duration estimation at 8 kHz sampling frequency.

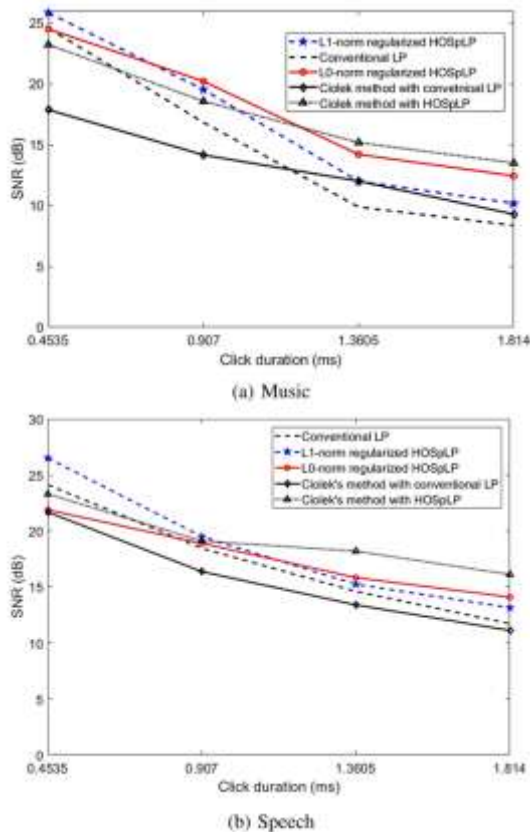


Figure 4: SNR of restored audio by using detection and restoration without any a priori knowledge on location and duration of click degradation.

Figure 5 show the SNR improvement obtained by using the backward prediction method with HOSpLP coefficients and Ciolek's method for the detection and restoration of click degraded audio sampled at 44.1 kHz. This is the difference between the SNR of the restored audio and the SNR of the click-degraded audio.

It is seen that all restoration methods achieve significant SNR improvement over the click-degraded audio. The proposed backward prediction method with HOSpLP coefficients for click detection and restoration is observed to lead to SNR improvement up to 4.5dB over Ciolek's method using conventional LP. On average both backward prediction and Ciolek's method performs similarly when using HOSpLP coefficients. This seems to indicate that the use of HOSpLP coefficients in both approaches is the reason for the improvement in restoration performance.

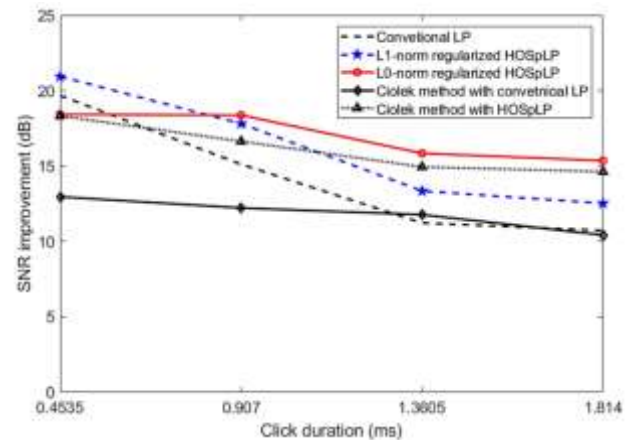


Figure 5: SNR improvement by detection and restoration without any a priori knowledge on location and duration of click degradation.

To see the impact of the sampling frequency on the restoration performance of the backward prediction method with HOSpLP and Ciolek's method, similar experiments were conducted for audio sampled at 8 kHz. Figure 6 shows the results of the detection and restoration for audio sampled at 8 kHz. At this sampling frequency the backward prediction method with HOSpLP coefficients leads to higher SNR for most click durations. The use of HOSpLP coefficients in Ciolek's method is observed to lead to better SNR as compared to conventional LP for higher click durations. This also shows the superior

### C. Perceptual evaluation of audio quality

PEAQ was used to estimate the subjective quality of the audio signal that is restored by using the proposed backward prediction method with HOSpLP coefficients and Ciolek's method. The PEAQ was calculated for each audio fragment as the original clean signal is available.

The result of each fragment was then averaged for each type of audio. Table III and III show the PEAQ evaluation obtained for by using the backward prediction method with HOSpLP, Ciolek's method and Ciolek's method with HOSpLP for music and speech respectively.

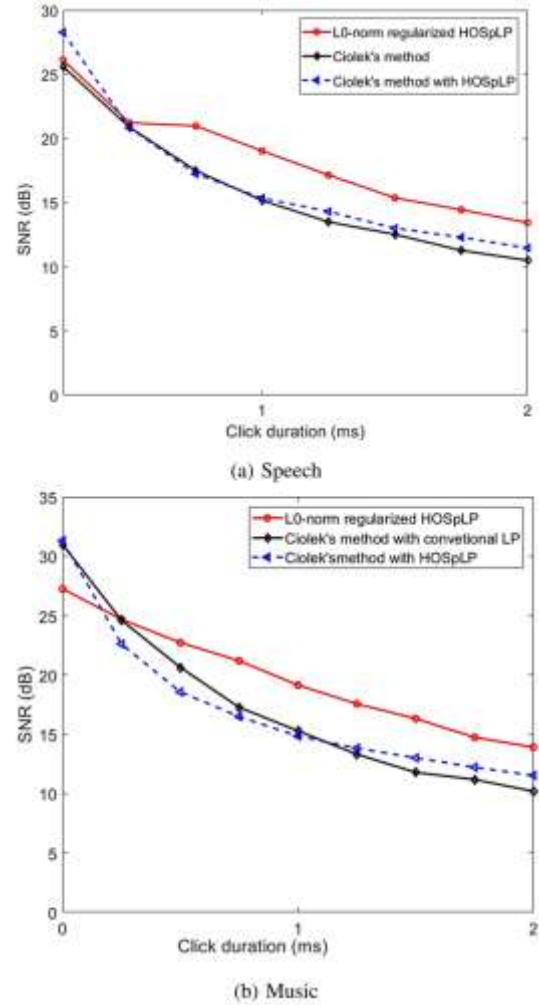


Figure 6: SNR of restored audio sampled at 8 kHz by using detection and restoration without any a priori on location and duration of click degradation.

It is seen that, the use of  $l_0$ -norm and  $l_1$ -norm regularized HOSpLP coefficients in the backward prediction click detection and then restoration leads to better PEAQ results as compared to conventional LP. However, it is noted that the  $l_1$ -norm regularized HOSpLP coefficients lead to higher PEAQ results as compared to  $l_0$ -norm regularized HOSpLP coefficients even though in terms of SNR  $l_0$ -norm regularized HOSpLP coefficients lead to better results. This may be attributed to the better modeling capabilities of  $l_1$ -norm regularized HOSpLP



coefficients especially for music. For speech, the use of HOSpLP coefficients in Ciolek's method is not observed to lead to significant improvement in PEAQ as compared to conventional LP Ciolek's method. However, for music the use of Table II: PEAQ evaluation for Music

HOSpLP coefficients in Ciolek's method leads to significant improvement in PEAQ as compared to conventional LP. This again, shows the better modeling capability of HOSpLP coefficients for music.

Method	Click duration in ms				
	0.454	0.907	1.361	1.814	2.268
Backward prediction with Conventional LP	-0.84	-1.13	-0.98	-1.29	-1.55
Backward prediction with $l1$ -norm HOSpLP	-0.67	-0.99	-0.85	-1.24	-1.34
Backward prediction with $l0$ -norm HOSpLP	-0.68	-0.81	-0.97	-1.27	-1.47
Ciolek's method	-1.13	-1.14	-0.93	-0.90	-0.95
Ciolek's method with $l1$ -norm HOSpLP	-0.65	-0.75	-0.62	-0.68	-0.91

Table III: PEAQ evaluation for speech

Method	Click duration in ms				
	0.454	0.907	1.361	1.814	2.268
Backward prediction with Conventional LP	-0.54	-0.64	-0.76	-0.79	-0.89
Backward prediction with $l1$ -norm HOSpLP	-0.37	-0.65	-0.75	-0.60	-0.77
Backward prediction with $l0$ -norm HOSpLP	-0.44	-0.56	-0.68	-0.76	-0.81
Ciolek's method	-0.67	-0.41	-0.49	-0.57	-0.59
Ciolek's method with $l1$ -norm HOSpLP	-0.38	-0.47	-0.46	-0.49	-0.54

#### D. Impact of amplitude of click degradation

A challenge for the click detection that has not been discussed is the amplitude of the click degradation, represented here by the variance of the assumed click generating-random process,  $\sigma_c^2$ . As the causes of click degradation are very diverse it is quite difficult to assume a single value for the variance of the click-generating random process. As such, even in a single recording, click degradation with very different amplitudes will be present. To evaluate the performance of the proposed HOSpLP-based click detection and restoration method for click degradations of different variance, the SNR improvement is evaluated by degrading the audio with click degradations having variance the same as the audio signal ( $\sigma_c^2 = \sigma_s^2$ ) and quarter of the audio signal ( $\sigma_c^2 = \frac{\sigma_s^2}{4}$ ).

Figure 7 shows the SNR improvement by the backward prediction method with HOSpLP and Ciolek's method with HOSpLP when the variance of the click generating random process is varied for speech and audio sampled at 8 kHz. It is seen that the three methods achieve significant SNR improvement. For click durations more than 0.5 ms, the proposed backward prediction method with HOSpLP and Ciolek's method with HOSpLP lead to a much better SNR improvement as the variance of the click-generating process decreases. However, for very short click durations, the backward prediction method with HOSpLP is inferior to Ciolek's method. It is also noted that as the variance of the click-generating random process decreases, Ciolek's method with HOSpLP leads to significant improvement as compared to the other two.



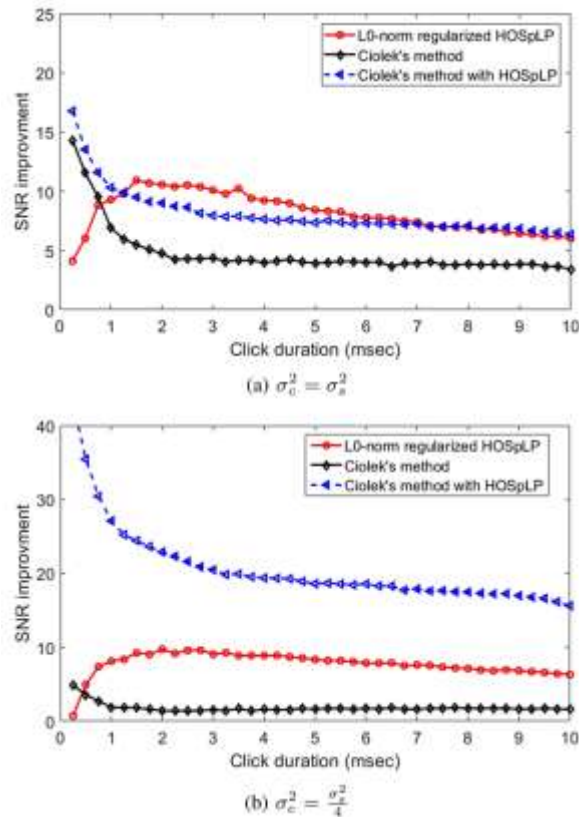


Figure 7: SNR improvement by detection and restoration without any a priori knowledge on location and duration of click degradation for music for different click degradation variance.

## CONCLUSIONS

In this paper, the use of high-order sparse linear predictions proposed for the detection of clicks and restoration of audio corrupted by click degradation. The use of the HOSpLP coefficients is suitable for both speech and tonal audio without a prior knowledge about the type of signal or click degradation. Several experiments were conducted to assess the performance of the proposed method in terms of click detection, restoration performance and robustness to the degrading click variance. The proposed methods achieved an improvement in SNR over conventional LP and a recently proposed method that also jointly detects and restores click degraded audio for speech and music. Even though both  $l1$ -norm and

$l0$ -norm regularized HOSpLP-based methods are not real-time, by using efficient ADMM and proximal gradient algorithm, the computation time can be limited to 2-3 times the duration of the frame under consideration on current general purpose computer. Considering the application at hand is for the restoration of archived audio media, the computational time is not expected to be a significant limitation.

Only artificial click degradation was considered in our experiments. A next step is to evaluate the proposed methods under real-life click degradation conditions. However, as the click-degraded samples are first discarded before restoration, working with real click degradations will only affect the detection and not the restoration performance.

## REFERENCES

- [1] Godsil S J and Rayner, P. J. W. *Digital audio restoration: a statistical model based approach*. Springer, January 1998.
- [2] Van Waterschoot T. and Moonen, M. Moonen, "Comparison of linear prediction models for audio signals," *EURASIP J. Audio, Speech, Music Process.*, vol. 20(5), pp. 1644–1657, July 2008.
- [3] Ruandaigh J. O. and Fitzgerald, W. Fitzgerald, "Interpolation of missing samples for audio restoration," *IEEE Electronics Letters*, vol. 30(8), pp. 622–623, April 1994.
- [4] Niedzwiecki M. and Ciolek M., "Elimination of clicks from archive speech signals using sparse autoregressive modeling," *Proc. 21st European Signal Process. Conf.*

- (*EUSIPCO 112*), pp. 2615–2619, August 2012.
- [5] Niedzwiecki M., Cioek, M. and Cisowski, K. “Elimination of impulsive disturbances from stereo audio recordings using vector autoregressive modeling and variable-order Kalman filtering,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 23(6), pp. 970–981, June 2015.
- [6] Ciolek M. and Niedzwiecki M., “Detection of impulsive disturbances in archive audio signals,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (New Orleans, LA, USA), March 2017.
- [7] Giacobello D., Christensen M. G., Murthi M. N., S. H. Jensen, and M. Moonen, “Sparse linear prediction and its applications to speech processing,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 20(5), pp. 1644–1657, July 2012.
- [8] Janssen A., Veldhuis R., and Vries L., “Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes,” *IEEE Trans., Acoust., Speech, Signal Process.*, vol. 34(2), pp. 317–330, Apr. 1986.
- [9] Kabal P. and Ramachandran R. P., “Joint optimization of linear predictors in speech coders,” *IEEE Trans. On Acoust., Speech and Signal Processing*, vol. 37(5), p. 642–650, May 1989.
- [10] Shi L., Jensen J. R., and Christensen M. G., “Least 1-norm pole zero modeling with sparse deconvolution for speech analysis,” in *Proc. 2017 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 17)*, (New Orleans, LA, USA), June 2017.
- [11] Giacobello D., Christensen M. Dahl G., J., Jensen S. H., and Moonen M., “Joint estimation of short-term and long-term predictors in speech coders,” in *Proc. 2009 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, (Taipei, Taiwan), pp. 409–412, IEEE, Apr. 2009.
- [12] Dufera B., Eneman D., K., and van Waterschoot T., “Missing sample estimation based on high-order sparse linear prediction for audio signals,” in *26th European Signal Processing Conference, EUSIPCO 2018*, (Roma, Italy), pp. 2464–2468, September 3–7, 2018.
- [13] Dufera B. D., Adugna E., Eneman K., and van Waterschoot T., “Restoration of click degraded speech and music based on high order sparse linear prediction,” in *IEEE AFRICON 2019*, (Accra, Ghana), September 25–27, 2019.
- [14] Giacobello D., van Waterschoot T., Christensen M. G., Jensen S. H., and Moonen M., “High-order sparse linear predictors for audio *Process. Conf. (EUSIPCO 110)*, (Aalborg, Denmark), pp. 234–238, August 2010.
- [15] Jensen T. L., Giacobello D., van Waterschoot T., and M. G. Christensen, “Fast algorithms for high-order sparse linear prediction with applications to speech processing,” *Speech Communication*, vol. 76(5), pp. 143–156, July 2016.
- [16] Hansen P. C., “Analysis of discrete ill-posed problems by means of the  $\ell_1$ -curve,” *SIAM Review*, vol. 34(4), pp. 561–580, Dec. 1992.

- [17] Toledano D. T., Gimenez A. O., Teixeira A., Rodriguez J. G., Gomez L. H., Hernandez R. S. S., and Castro D. R., “*Advances in Speech and Language Technologies for Iberian Languages*”. Springer, November 2012.
- [18] Antonello N., Stella L., Patrinos P., and van Waterschoot T., “Proximal gradient algorithms: Applications in signal processing,” *arXiv:1803.01621*, March 2018.
- [19] Vaseghi S. V. and Rayner P. J. W., “Detection and suppression of impulsive noise,” in *speech communication systems. IEE Proceedings*, pp. 38–46, 1990.
- [20] Niedzwiecki M. and Ciolek M., “Renovation of archive audio recordings using sparse autoregressive modeling and bidirectional processing,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (Vancouver, BC, Canada), May 2013.
- [21] Niedzwiecki M. and Cisowski K., “Adaptive scheme for elimination of broadband noise and impulsive disturbances from ar and arma signals,” *IEEE Trans. on Audio, Speech, Lang. Processing*, vol. 14(1), pp. 967–982, March 1996.
- [22] Ciolek M. and Niedzwiecki M., <http://eti.pg.edu.pl/katedra-systemow-automatyki/ICASSP2017>.
- [23] Bang and Olufsen, “Music for archimedes,”
- [24] Avila F. R. and Biscainho L. W. P., “Bayesian restoration of audio signals degraded by impulsive noise modeled as individual pulses,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 20(9), pp. 2470–2480, November 2012.
- [25] ITU-R, “Method for objective measurements of perceived audio quality,” Recommendation 1387-1, International Telecommunication Union, 1998-2001.
- [26] Kabal P., “An examination and interpretation of itu-r bs.1387: Perceptual evaluation of audio quality,” tsp lab technical report, Dept. Electrical and Computer Engineering, McGill University, May 2002.

## NATIONAL AND INTERNATIONAL ADVISORY BOARD

Prof. Abrham Engida, Michigan State University, USA  
Ato Asrat Bulbula, Consultant, Ethiopia  
Dr. Anuradha Jabasingh, Addis Ababa University  
Dr. Beshawired Ayalew, Clenson University, USA  
Prof. Carlo Rafele, Politecnico, Italy  
Prof. Ja Choon Koo, Sungkyunkwan University, Korea  
Prof. Amde M. Amde, University of Maryland, USA  
Prof. Beyong Soo Lim National University of Korea  
Dr. Fekadu Shewarega, Universitaet-Duisburg, Essen, Germany  
Prof. Gunter Busch, TU-Cottbus, Cottbus, Germany  
Dr. Kibret Mequanint, University of Western Ontario, Canada  
Dr. Mekonnen Gebremichael, University of Connecticut, USA  
Dr. Mulugeta Metaferia, Consultant, Ethiopia  
Dr. Solomon Assefa, IBM, USA  
Dr. Tesfaye Bayou, Consultant, Ethiopia  
Dr. Woubshet Berhanu, Self Help Africa, Ethiopia

## ACKNOWLEDGEMENTS

The Editorial Board of Zede Journal of Ethiopian Engineers and Architects would like to express its sincere gratitude to the following individuals for reviewing the manuscripts that were originally submitted for publication in Zede Volume: 40

Prof. Abebe Dinku	Ing. Getaneh Terefe	Berhanu Bekeko (Managing Editor)
Prof. Girma Z/Yohannes	Dr. Geremew Sahilu	
Prof. Ngendra P. Singh	Dr. Getachew Alemu	
Dr. Demis Alemu	Dr. Murad Rediwan	
Dr. Agizew Nigussie	Dr. Mengesha Mamo	
Dr. Beneyam Berhanu	Dr. Celestin Nkundineza	
Dr. Birouktawit Taye	Dr. Abrham Assefa	
Dr. Heyaw Terefe	Dr. Kassahun Admasu	