Athens Journal of Sciences

Quarterly Academic Periodical, Volume 11, Issue 2, June 2024

URL: https://www.athensjournals.gr/ajs

Email: journals@atiner.gr

e-ISSN: 2241-8466 DOI: 10.30958/ajs

Front Pages

SEMRA BENZER & RECEP BENZER

The Growth of Crayfish, which Serves as an Indicator of Clean and Healthy Water Ecosystems in the Mediterranean Region

EVANGELOS XEVELONAKIS & TANBIR MANN

Identifying Sustainability Efforts in Company's Reports Using Text Mining and Machine Learning

ILDA RUSI & ALBI ALLIAJ

Generative Urban Design in the Field of Infrastructure: An Optimizing Solution for Connecting Fier and Vlora County by a 600 m Bridge over Selenica River, Albania

CHIGOZIE NWEKE-EZE

Infrastructures of Large-Scale Geothermal Energy Projects in Kenya: Materialization, Generativity, and Socio-Economic Development Linkages

Athens Journal of Sciences

Published by the Athens Institute for Education and Research (ATINER)

****

ATINER is an Athens-based World Association of Academics and Researchers based in Athens. ATINER is an independent and non-profit Association with a Mission to become a forum where Academics and Researchers from all over the world can meet in Athens, exchange ideas on their research and discuss future developments in their disciplines, as well as engage with professionals from other fields. Athens was chosen because of its long history of academic gatherings, which go back thousands of years to Plato's Academy and Aristotle's Lyceum. Both these historic places are within walking distance from ATINER's downtown offices. Since antiquity, Athens was an open city. In the words of Pericles, Athens"…is open to the world, we never expel a foreigner from learning or seeing". ("Pericles' Funeral Oration", in Thucydides, The History of the Peloponnesian War). It is ATINER's mission to revive the glory of Ancient Athens by inviting the World Academic Community to the city, to learn from each other in an environment of freedom and respect for other people's opinions and beliefs. After all, the free expression of one's opinion formed the basis for the development of democracy, and Athens was its cradle. As it turned out, the Golden Age of Athens was in fact, the Golden Age of the Western Civilization. Education and (Re)searching for the 'truth' are the pillars of any free (democratic) society. This is the reason why Education and Research are the two core words in ATINER's name.

****

The ***Athens Journal of Sciences (AJS)*** is an Open Access quarterly double-blind peer reviewed journal and considers papers from all areas of Natural & Formal Sciences, including papers on agriculture, computer science, environmental science, materials science, transportation science, chemistry, physics, mathematics and statistics, biology, geography, and earth science (geology, oceanography, astronomy, meteorology). Many of the papers published in this journal have been presented at the various conferences sponsored by the Natural & Formal Sciences Division of the Athens Institute for Education and Research (ATINER). All papers are subject to ATINER's Publication Ethical Policy and Statement.

The Athens Journal of Sciences
ISSN NUMBER: 2241-8466- DOI: 10.30958/ajs
Volume 11, Issue 2, June 2024
Download the entire issue (PDF)

Front Pages i-viii

The Growth of Crayfish, which Serves as an Indicator of Clean and Healthy Water Ecosystems in the Mediterranean Region 75
Semra Benzer & Recep Benzer

Identifying Sustainability Efforts in Company's Reports Using Text Mining and Machine Learning 89
Evangelos Xevelonakis & Tanbir Mann

Generative Urban Design in the Field of Infrastructure: An Optimizing Solution for Connecting Fier and Vlora County by a 600 m Bridge over Selenica River, Albania 109
Ilda Rusi & Albi Alliaj

Infrastructures of Large-Scale Geothermal Energy Projects in Kenya: Materialization, Generativity, and Socio-Economic Development Linkages 125
Chigozie Nweke-Eze

Athens Journal of Sciences
Editorial and Reviewers' Board

President's Message

All ATINER's publications including its e-journals are open access without any costs (submission, processing, publishing, open access paid by authors, open access paid by readers etc.) and is independent of presentations at any of the many small events (conferences, symposiums, forums, colloquiums, courses, roundtable discussions) organized by ATINER throughout the year and entail significant costs of participating. The intellectual property rights of the submitting papers remain with the author. Before you submit, please make sure your paper meets the basic academic standards, which includes proper English. Some articles will be selected from the numerous papers that have been presented at the various annual international academic conferences organized by the different divisions and units of the Athens Institute for Education and Research. The plethora of papers presented every year will enable the editorial board of each journal to select the best, and in so doing produce a top-quality academic journal. In addition to papers presented, ATINER will encourage the independent submission of papers to be evaluated for publication.

The current issue is the second of the eleventh volume of the *Athens Journal of Sciences (AJS),* published by Natural & Formal Sciences Division of ATINER.

Gregory T. Papanikos, President, ATINER.

Athens Institute for Education and Research
A World Association of Academics and Researchers

12th Annual International Conference on Chemistry
22-25 July 2024, Athens, Greece

The Natural Sciences Unit of ATINER, will hold its **12th Annual International Conference on Chemistry, 22-25 July 2024, Athens, Greece** sponsored by the Athens Journal of Sciences. The aim of the conference is to bring together academics and researchers of all areas of chemistry and other related disciplines. You may participate as stream organizer, presenter of one paper, chair a session or observer. Please submit a proposal using the form available (https://www.atiner.gr/2023/FORM-CHE.doc).

Academic Members Responsible for the Conference

- **Dr. Haiduke Sarafian**, Head, Natural Sciences Unit, ATINER & Professor of Physics and Endowed Chair of John T. and Paige S. Smith Professor of Science, Pennsylvania State University, USA.

Important Dates

- Abstract Submission: **11 June 2024**
- Acceptance of Abstract: 4 Weeks after Submission
- Submission of Paper: **24 June 2024**

Social and Educational Program

The Social Program Emphasizes the Educational Aspect of the Academic Meetings of Atiner.

- Greek Night Entertainment (This is the official dinner of the conference)
- Athens Sightseeing: Old and New-An Educational Urban Walk
- Social Dinner
- Mycenae Visit
- Exploration of the Aegean Islands
- Delphi Visit
- Ancient Corinth and Cape Sounion

Conference Fees

Conference fees vary from 400€ to 2000€
Details can be found at: https://www.atiner.gr/fees

Athens Institute for Education and Research
A World Association of Academics and Researchers

12th Annual International Conference on Physics
22-25 July 2024, Athens, Greece

The Natural Sciences Unit of ATINER, will hold its **12th Annual International Conference on Physics, 22-25 July 2024, Athens, Greece** sponsored by the Athens Journal of Sciences. The aim of the conference is to bring together academics and researchers of all areas of physics and other related disciplines. Please submit a proposal using the form available (https://www.atiner.gr/2023/FORM-PHY.doc).

Important Dates

- Abstract Submission: **11 June 2024**
- Acceptance of Abstract: 4 Weeks after Submission
- Submission of Paper: **24 June 2024**

Academic Member Responsible for the Conference

- **Dr. Haiduke Sarafian**, Head, Natural Sciences Unit, ATINER & Professor of Physics and Endowed Chair of John T. and Paige S. Smith Professor of Science, Pennsylvania State University, USA.

Social and Educational Program

The Social Program Emphasizes the Educational Aspect of the Academic Meetings of Atiner.

- Greek Night Entertainment (This is the official dinner of the conference)
- Athens Sightseeing: Old and New-An Educational Urban Walk
- Social Dinner
- Mycenae Visit
- Exploration of the Aegean Islands
- Delphi Visit
- Ancient Corinth and Cape Sounion

More information can be found here: https://www.atiner.gr/social-program

Conference Fees

Conference fees vary from 400€ to 2000€
Details can be found at: https://www.atiner.gr/fees

Athens Journal of Sciences- Volume 11, Issue 2, June 2024 – Pages 75-88

The Growth of Crayfish, which Serves as an Indicator of Clean and Healthy Water Ecosystems in the Mediterranean Region

By Semra Benzer[*] & Recep Benzer[±]

The narrow-fingered freshwater crayfish, also known as Pontastacus leptodactylus in scientific terms and referred to as such in Turkish, is a prevalent species in the inland waters of the Mediterranean region. This crayfish serves as an indicator of clean and healthy water ecosystems due to its high sensitivity to pollution and environmental changes. Its presence signals the maintenance of water quality and a balanced ecosystem. In a study on 283 freshwater crayfish in Hirfanlı Dam Lake during July and August 2023, it was observed that 53.36% were male and 46.64% were female. Length ranged from 80.44 mm to 121.11 mm, and weight varied between 11.61 g and 43.93 g. Average length and weight for males were 95.9874 ± 6.4603 mm and 24.1861 ± 5.3696 g, for females were 95.0981 ± 5.5519 mm and 21.5193 ± 3.4488 g, and for combined sexes were 95.5724 ± 6.0594 mm and 22.9422 ± 4.7579 g. The length-weight relationship (LWR) was established for females ($W=0.00166359 \times TL^{2.0767}$), males ($W=0.00019587\times TL^{2.5651}$), and all individuals ($W=0.00037111 \times TL^{2.4162}$). Exponential values "b" for LWR were 2.0767 (r^2=0.993), 2.5651 (r^2=0.986), and 2.4162 (r^2=0.987), respectively. The study compared traditional LWR approaches with artificial intelligence methods for growth analysis, suggesting the latter as a viable alternative in assessing freshwater crayfish growth in aquatic systems.

Keywords: *length-weight relationships, crayfish, Mediterranean region, inland water, artificial intelligent*

Introduction

Freshwater crayfish, commonly known as crayfish (*Astacus leptodactylus* Eschscholtz, 1823), stands out as a species of considerable ecological and economic significance on a global scale. Within our nation (Turkey), it represents a species sourced from inland waters, featured among export commodities (Kozák et al. 2015, Yazıcıoğlu et al. 2018). Crayfish are invertebrates distributed worldwide, excluding Antarctica and the African continent, with nearly 540 species (Madrigal-Bujaidar et al. 2017). Particularly known for their potential in food consumption, the narrow-clawed crayfish (*Astacus leptodactylus* Eschscholtz 1823), is one of the well-recognized and valuable European freshwater species, residing in a wide range of freshwater habitats and estuaries (Köksal 1988).

The freshwater crayfish are acknowledged as effective bioindicators for assessing the health of aquatic ecosystems. These organisms play a significant role

[*]Professor, Gazi University, Türkiye.
[±]Associate Professor, Konya Food and Agriculture University, Türkiye.

in aquatic food webs, attaining substantial biomass as macroinvertebrates. Moreover, they prove to be valuable bioindicators in determining water quality and responding to physiological and contaminant stressors (Garabaghi et al. 2022).

Koutrakis et al. (2007) state, based on genetic and morphological data of crayfish species such as *Austropotamobius torrentium*, *Astacus astacus*, and *Astacus leptodactylus*, that the region is rich in biodiversity and provides a suitable habitat for these species in Greece. *Astacus leptodactylus* is the only significant freshwater crayfish species in Turkey, and it is naturally and widely observed in lakes, ponds, and streams throughout the country (Gören and Karayücel 2022). In their study presenting the global systematic list of freshwater crayfish, Crandall and De Grave (2017) specified *Pontastacus leptodactylus* Eschscholtz, 1823, as a synonym of *Astacus leptodactylus* Esch., 1823.

Fisheries research and management frequently employ biometric relationships as a method to convert field-collected data into appropriate indices (Anderson and Gutreuter 1983, Roul et al. 2020). However, when it comes to analyzing fisheries data, the most widely utilized tool is the length-weight relationships (LWRs) (Türker et al. 2018). LWRs are instrumental in calculating body condition indices and predicting weight based on a known length (Froese et al. 2011, Dash et al. 2023).

Studies encompassing morphometric analyses and evaluations on *Astacus leptodactylus/Pontastacus leptodactylus* have been identified as (Benzer and Benzer 2018, Berber et al. 2020, Gören and Karayücel 2022, Garabaghi et al. 2022, Benzer and Benzer 2022, Gültepe et al. 2023, Boyalık et al. 2023, Dartay 2023, Alvanou et al. 2024, Roljić et al. 2024).

The key variables in modeling the growth of an aquatic organism are typically the length and weight. Fundamentally, a length-weight relationship model is derived through linear regression, which is employed to calculate the correlation between the growth in length and the growth in weight (Munro and Pauly 1983). These models play a crucial role in understanding the organism's development and assessing populations in aquatic ecosystems. During underwater visual censuses, transforming length data into weight data is common, allowing for biomass estimation and monitoring organism populations in ecosystems (Samoilys 1997). However, the relationship between length and weight is often non-linear (Froese 2006), and transformations using linear regression methods may lead to low predictive values. In such cases, traditional statistical analysis methods, especially single or multiple linear regression models, may be limited in terms of quantification and prediction (Suryanarayana et al. 2008, Türeli Bilen et al. 2011, Benzer et al. 2017, Benzer and Benzer 2019, Benzer and Benzer 2022).

Currently, the application of Artificial Neural Networks (ANNs) in predictive modeling presents a promising alternative to traditional statistical approaches, particularly in cases where non-linear patterns exist. Artificial neural networks are computer algorithms that simulate the activity of neurons and information processing in the human brain (Zou et al. 2009). Unlike more commonly used regression models, neural networks do not require a specific functional relationship or distribution assumptions about the data, eliminating the need for data transformation. This characteristic makes neural network modeling a powerful tool for exploring complex, non-linear biological problems, such as those encountered in fisheries research (Suryanarayana et al. 2008).

Artificial neural networks bring flexibility and a comprehensive approach, enabling more effective resolution of non-linear relationships and complexities than conventional statistical models (Ezziane 2006, Benzer and Benzer 2023a). Consequently, artificial neural networks provide an efficient means of understanding and predicting intricate biological processes, such as the growth and development of organisms in aquatic ecosystems. Therefore, artificial neural networks emerge as a more reliable and effective tool in situations where non-linear patterns prevail than traditional methods.

There have been numerous studies on predictive modeling using neural networks for the growth of aquatic organisms, specifically focusing on fish (Benzer and Benzer 2016, Benzer and Benzer 2017, Özcan and Serdar 2018, Özcan and Serdar 2019, Özcan 2019, Benzer and Benzer 2020, Benzer and Benzer 2023b, Akkan et al. 2024) and crayfish (Türeli Bilen et al. 2011, Benzer et al. 2015, Benzer and Benzer 2018, Benzer and Benzer 2022).

The main purpose of this study is to assess the length-weight measurements of crayfish supplied from the ecosystem using both traditional methods and the Artificial Neural Networks approach, a sub-branch of artificial intelligence.

Methodology

As part of a project conducted by Gazi University, a classification study will be carried out to determine the genders of crayfish in different regions. However, the focal point of this study is the traditional and artificial intelligence-assisted assessment of size and weight, specifically on crayfish samples collected from Hirfanlı Dam Lake. Hirfanlı Dam, located between Şereflikoçhisar in Kırşehir province, was constructed between 1953 and 1959 for the purposes of energy production and flood control. Additionally, the villages surrounding Hirfanlı Reservoir within the borders of Ankara, Kırşehir, and Aksaray provinces cover an area of 26,300 hectares, making fishing a significant source of income (DSI 1968).

Local fishermen engaged in commercial fishing with fyke nets captured crayfish during the 2023 fishing seasons. A total of 283 crayfish, consisting of 132 females and 151 males, were scrutinized in the study. The captured crayfish were identified, transported to the laboratory, and subjected to the necessary metric measurements to determine their gender. Measurements, including carapace length (CL), abdomen width (Aw), abdomen length (AL), carapace width (Cw), chela length (ChL), chela width (Chw), total length (TL), and total weight (TW), were taken using a caliper with a sensitivity of 0.5 mm, and their weight was measured with a scale having a sensitivity of 0.01 g.

For most crayfish species, the connection between length (L) and weight (W) can be effectively captured by the "length-weight relationship" equation (Ricker, 1973):

$$W = a\,L^b$$

In this equation, *W* represents the weight of the crayfish in grams, *L* signifies the length in centimeters, and the variables '*a*' and '*b*' denote constants that define the relationship between length and weight.

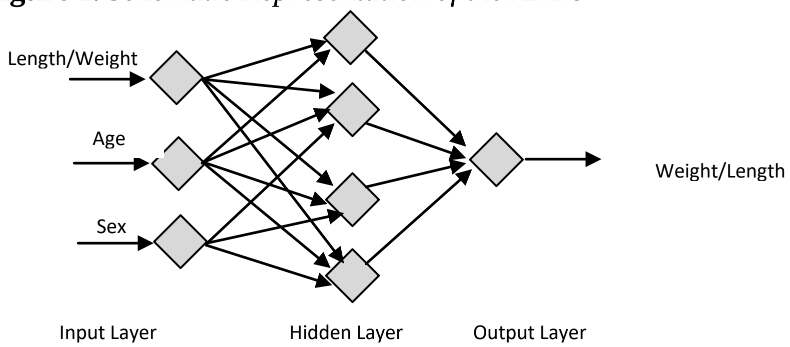
Taking cues from the complex data analysis and learning abilities exhibited by the human brain, ANNs have been proposed to streamline the intricate process of analyzing complex data and, in the end, arriving at well-informed decisions by simulating the neural system found in the human brain. In this study, the Back-Propagation Networks, a supervised learning method with a specific network structure, will address problem-solving, and the data evaluation will be conducted using the WEKA application.

Generally, ANNs consist of three primary layers, commencing with the input layer and concluding with the output layer, which is intricately linked to one or more hidden layers where data undergoes intricate processing. The activation function or transfer function in ANNs commonly takes the form of a sigmoid, although there is room for variation (Hopfield 1988). For this study, a well-regarded ANNs model known as the multilayer perceptron has been devised, featuring a hidden layer with five nodes and operating with a learning rate of 0.03 and momentum of 0.2 (Figure 1).

Metrics play a pivotal role in monitoring and measuring the model's performance during both training and testing phases. The Mean Absolute Percentage Error (MAPE) serves as an error measure, with a lower result indicative of higher performance, inversely proportional to performance (Wang and Xu 2004). Mathematically, the performance metric within ANNs can be expressed as:

$$MAPE = \frac{100}{n} \sum_j \frac{|e_j|}{|A_j|}$$

Figure 1. *Schematic Representation of the ANNs*



Results and Discussion

The investigation identified that out of the total 283 crayfish captured, 132 individuals (46.64%) were identified as females, while 151 (53.36%) were males, resulting in a female-to-male ratio of 0.87/1.00. The metric lengths of crayfish, including TL, CL, AL, Aw, ChL, Chw, and Cw, displayed variations spanning from 80.44 to 121.11, 38.14 - 62.26, 40.22 - 65.09, 16.57 – 39.49, 45.63 – 99.91, 6.88 - 19.89, 18.08 - 75.50 mm, respectively. Additionally, crayfish's weight fluctuated between 11.61 and 43.93 g, with an average weight of 22.92 g (Table 1).

Table 1. *Metric Parameters for Crayfish*

Parameter	Sex	Average ± S_x	Min-Max	t test
TL	♀ ♂ ♀♂	95.098 ± 0.483 94.540 ± 0.525 95.570 ± 0.360	80.44 – 112.38 82.00 – 121.11 80.44 – 121.11	p<0.05
TW	♀ ♂ ♀♂	21.466 ± 0.293 22.780 ± 0.436 22.910 ± 0.281	14.92 - 32.94 11.61 - 43.93 11.61 - 43.93	p<0.05
CL	♀ ♂ ♀♂	46.139 ± 0.273 47.600 ± 0.289 47.260 ± 0.209	38.14 - 57.17 40.74 - 62.26 38.14 - 62.26	p<0.05
AL	♀ ♂ ♀♂	48.958 ± 0.269 46.850 ± 0.307 48.300 ± 0.209	40.22 – 59.27 41.26 – 65.09 40.22 – 65.09	p<0.05
Aw	♀ ♂ ♀♂	21.283 ± 0.156 19.840 ± 0.208 20.750 ± 0.136	17.30 - 27.27 16.57 - 39.49 16.57 - 39.49	p<0.05
ChL	♀ ♂ ♀♂	56.544 ± 0.429 65.910 ± 0.721 63.230 ± 0.575	45.63 – 69.21 49.17 – 99.91 45.63 - 99,91	p<0.05
Chw	♀ ♂ ♀♂	10.612 ± 0,140 11.540 ± 0,200 11.410 ± 0,135	6.88 - 15.68 7.50 - 19.89 6.88 - 19.89	p<0.05
Cw	♀ ♂ ♀♂	24.653 ± 0.401 24.760 ± 0.518 25.160 ± 0,334	20.20 - 71.96 18.08 - 75.50 18.08 - 75.50	p<0.05

S_x = Standard error

The TL, AL, and Aw measurements were higher in females than males, whereas TW, CL, ChL, Chw, and Cw were lower in females than males (Table 1). However, the differences between females and males in TL, TW, CL, AL, Aw, ChL, Chw, and Cw were statistically insignificant ($p < 0.05$).

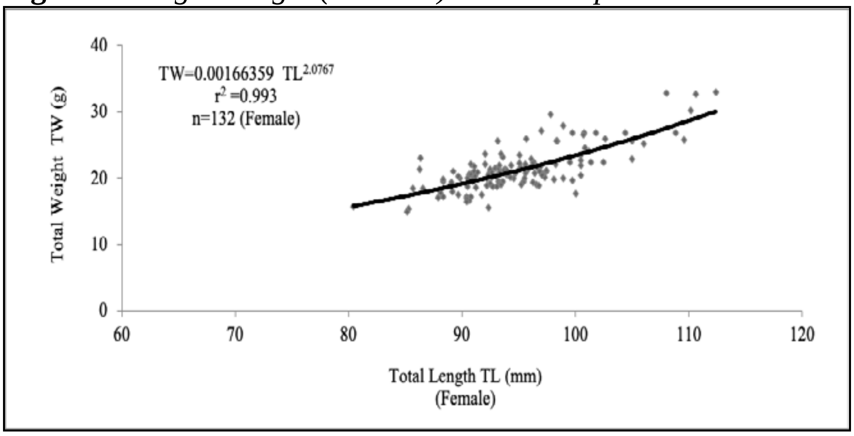
Table 2. *LWR Parameters, Equations and Correlation Coefficients*

Species	Sex	Relationship	R^2
TL – TW	♀ ♂ ♀♂	$TW = 0.00166359 \times TL^{2.0767}$ $TW = 0.00019587 \times TL^{2.5651}$ $TW = 0.00037111 \times TL^{2.4162}$	0.993 0.986 0.987

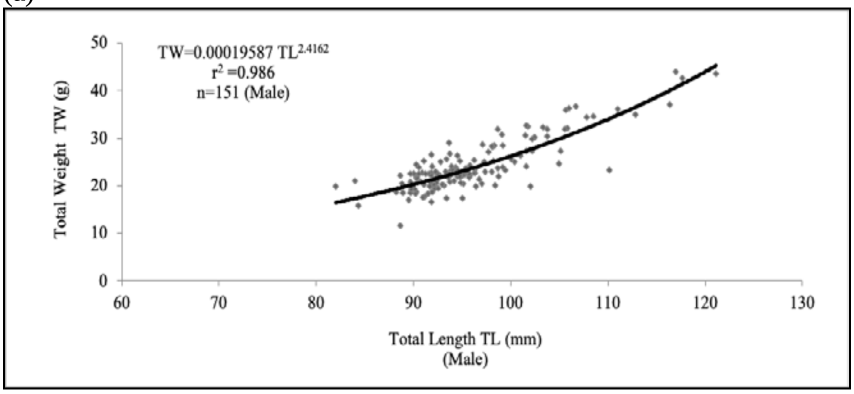
The analysis of the crayfish revealed distinct TL – TW relationships across genders: for females (Figure 2a), $TW = 0.00166359 \times TL^{2.0767}$ ($R^2 = 0.993$), for males (Figure 2b), $TW = 0.00019587 \times TL^{2.5651}$ ($R^2 = 0.986$), and for all genders (Figure 2c), $TW = 0.00037111 \times TL^{2.4162}$ ($R^2 = 0.987$). Figure 2 provides separate graphical representations of these relationships for female, male, and all individuals, while the distribution of male, female, and all individuals can be observed in a single figure (Figure 2d). Additionally, the TL-TW data underwent artificial neural network evaluation using MATLAB, and are depicted in Figure 3.

In this investigation involving crayfish specimens, TL-TW values were analyzed across three distinct categories: Crayfish data, regression data, and ANNs data. The resultant mean absolute percentage errors (MAPE) are detailed in Table 3. Furthermore, it's noteworthy that mean absolute error serves as a prominent loss function, particularly in trend estimation within statistical analyses, representing a crucial metric for assessing the prediction accuracy of an estimation method.

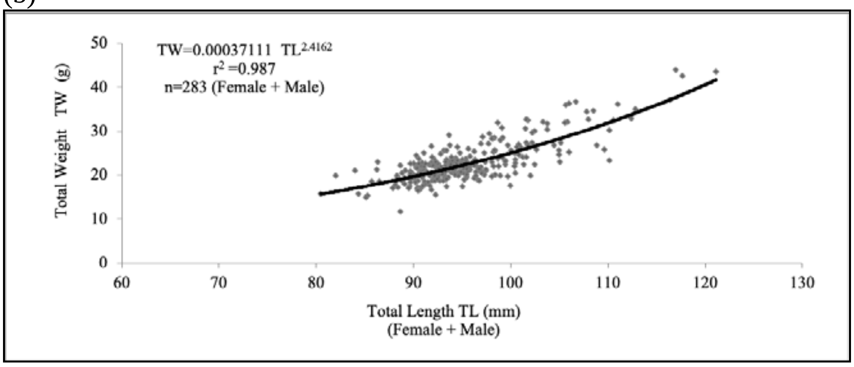
Figure 2. *Length-Weight (TL – TW) Relationships*



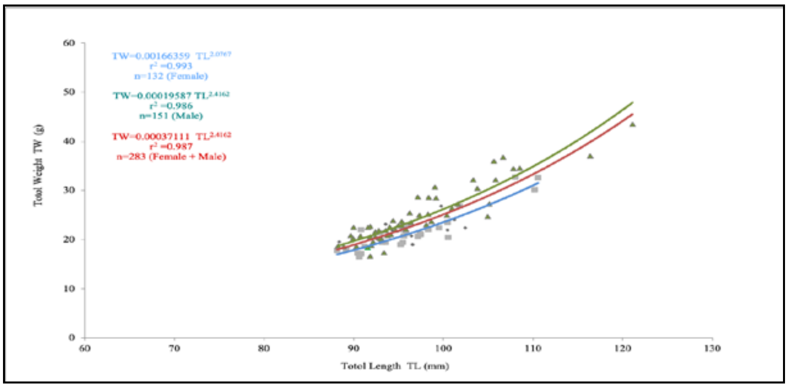
(a)



(b)



(c)



(d)

Figure 3. *Length-Weight (TL – TW) Relationships (ANNs)*

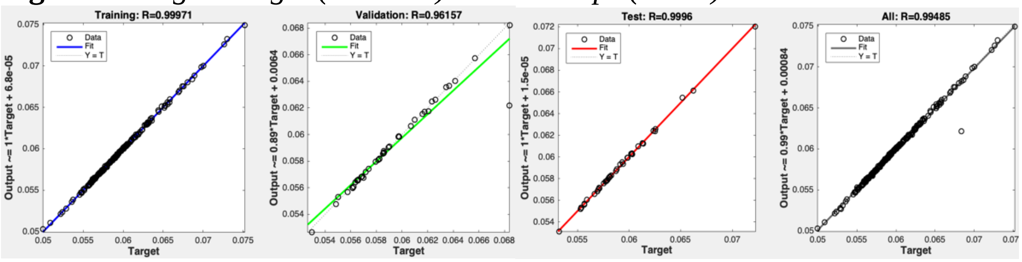


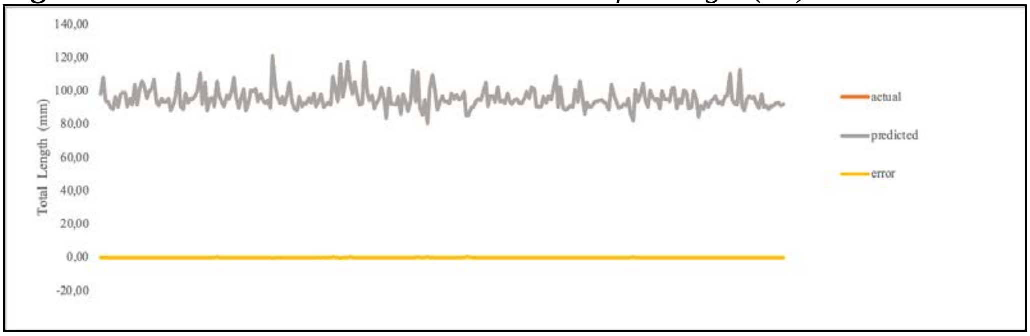
Table 3. *ANNs and LWR Values with MAPE*

	LWRs				ANNs			
	LWRs		MAPE		ANNs		MAPE	
	L	W	L	W	L	W	L	W
♀	21.3331	95.4969	0.62	0.42	20.82	95.10	3.09	0.004
♂	23.7932	96.6017	1.65	0.64	21.66	95.99	11.68	0.003
♀♂	22.6152	96.1423	1.34	0.59	21.27	95.58	7.76	0.003

When considering the ANNs' MAPE values across females, males, and all individuals compared to the LWRs' MAPE values, it is evident that the former are lower, indicating the accuracy and effectiveness of the ANNs' predictions. This suggests that the artificial neural networks perform quite well in estimating the length-weight relationships. However, while not as pronounced as in length predictions, the performance in weight predictions is also notably good (Table 3).

Several factors may contribute to these outcomes. Firstly, the ANN model's complexity and the training data's quality may enhance accuracy. Additionally, a balanced dataset and appropriate model tuning play significant roles. Lastly, the comparison between LWRs' and ANNs' predictions reveals the superior precision and accuracy of the ANNs' model. This underscores the superiority of artificial neural networks over traditional methods in LWRs predictions. These findings emphasize the importance of utilizing and refining artificial neural networks in future research, highlighting their potential to provide a broader accuracy range in LWRs predictions.

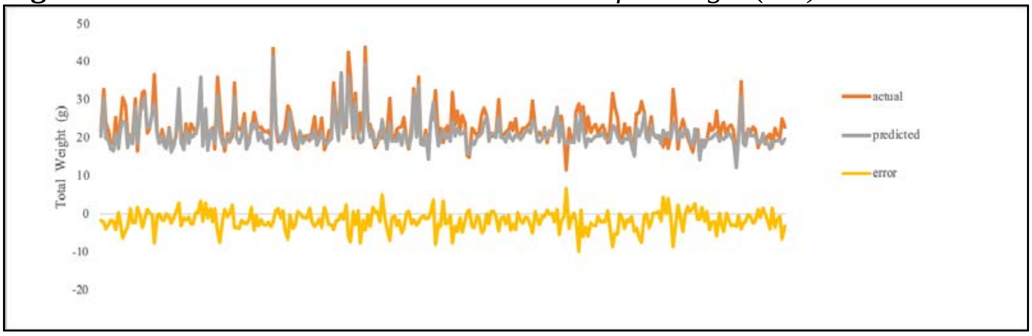
Figure 4. *Actual and Predicted Data with Errors for Length (TL)*



When we examine the comparison between actual and predicted values based on gender, it's evident that prediction errors vary. For instance, in the female group, errors range from -0.12 to 0.287, while in the male group, errors range from -0.638 to 0.272. This indicates variability in the accuracy of predictions across different samples (Figure 4).

Furthermore, while some predictions show small errors, others exhibit larger deviations. Overall, it's observed that errors are balanced between positive and negative values, indicating that predictions both overestimate and underestimate the true values. Further analysis could be conducted to explore the factors contributing to these differences, including the features used in prediction models, the complexity of algorithms, and the quality of training data. Adjustments in these aspects could improve the accuracy of future predictions.

Figure 5. *Actual and Predicted Data with Errors for Weight (TW)*



When examining the weight prediction data, it can be observed that the predicted weights differ from the actual values. In some cases, the predictions are either significantly higher or lower than the actual weights, while in other cases, they are closer. For instance, some predictions exhibit large discrepancies ranging from -7.722 to 5, while others show smaller variances between -0.10 and 4.447.

Further analysis of this data could help understand why some predictions are more accurate than others and identify factors influencing prediction accuracy. This deeper analysis could lead to the development of strategies to improve predictions.

Table 4. *ANNs Crayfish Results in Literature*

Location	Gender	MAPE (%) ANNs		MAPE (%) LWR		Reference
		L	W	L	W	
Mogan Lake (Ankara, Turkey)	♀	1.234	11.552	0.482	0.866	Benzer et al. 2015
	♂	9.763	7.114	1.237	1.713	
	♀♂	0.108	1.703	1.750	1.727	
Hirfanlı Dam Lake (Kırşehir, Turkey)	♀	12.36	4.97	1.53	4.29	Benzer et al. 2017
	♂	5.85	1.85	2.13	5.82	
	♀♂	8.39	7.41	2.33	6.63	
Eğirdir Lake (Isparta, Turkey)	♀	1.277	3.909	0.771	2.074	Benzer et al. 2017
	♂	0.140	1.206	1.365	3.725	
	♀♂	0.436	2.409	1.435	4.087	
Uluabat Lake (Bursa, Turkey)	♀	2.18	5.28	1.94	5.20	Benzer and Benzer 2018
	♂	2.21	5.75	2.22	6.28	
	♀♂	2.65	0.64	3.01	5.89	
Yeniçağa Lake (Bolu, Turkey)	♀	1.491	3.685	1.695	3.828	Benzer and Benzer 2020
	♂	1.768	4.710	2.391	5.098	
	♀♂	0.064	1.153	2.069	4.603	
İznik Lake (Bursa, Turkey)	♀	0.44	3.64	2.63	6.44	Benzer and Benzer 2022
	♂	0.26	2.79	3.75	9.22	
	♀♂	0.09	4.35	3.63	8.96	
Hirfanlı Dam Lake (Kırşehir, Turkey)	♀	3.09	0.004	0.62	0.42	This study
	♂	11.68	0.003	1.65	0.64	
	♀♂	7.76	0.003	1.34	0.59	

When the results of this study are compared with the literature (Table 4), it is observed that the performance of ANNs in predicting freshwater crayfish populations is influenced by the differences in MAPE values reported in various studies conducted in different lakes and ponds in Turkey. There are notable differences in MAPE values among female (♀), male (♂), and combined (♀♂) crayfish populations. The MAPE values reported in this study are lower than those reported in the literature for the same locations and genders. This indicates that the ANNs used in this study may be more effective in predicting freshwater crayfish populations with lower error rates than previous approaches. Consistently, the ANNs in this study exhibit relatively low MAPE values compared to the literature across different locations and genders. This suggests a degree of reliability in the prediction capabilities of ANNs despite variations in environmental factors or crayfish populations. While the ANNs in this study generally perform well compared to previous approaches, there are still instances where MAPE values are relatively high, indicating potential for improvement. Future research could focus on further enhancing ANNs models or incorporating additional features to increase prediction accuracy.

Conclusions

In conclusion, this study suggests that ANNs can be considered as a significant alternative for predicting freshwater crayfish populations. The findings indicate that ANNs are effective in predicting freshwater crayfish populations, influenced by variations in MAPE values reported in various studies conducted in different lakes and ponds. This study emphasizes that ANNs emerge as a potential strong alternative for predicting freshwater crayfish populations with lower error rates than previous methods.

Overall, the ability of ANNs to predict freshwater crayfish populations could serve as a valuable tool in managing and conserving such ecosystems. The findings of this study encourage further research to enhance ANN models or improve prediction accuracy in future studies. These improvements could contribute significantly to monitoring and managing freshwater crayfish populations more effectively.

Additionally, in the future, a classification study will be conducted regarding freshwater crayfish, aiming to determine the genders of these particular freshwater crayfish using artificial intelligence algorithms. This study aims to contribute to our understanding of the ecological dynamics of these species and to develop comprehensive management strategies for freshwater ecosystems.

Acknowledgments

This study was financially supported by the Gazi University Project Resources Center Fund (BAP Project Number: SPD/2023-8461).

References

Akkan T, Mutlu C, Hakan IŞIK, Yazicioğlu O, Yazici R, Yilmaz M, et al. (2024) Determination of fish condition factor using artificial neural networks and machine learning algorithms. *Acta Biologica Turcica* 37(1): 8–1.

Alvanou MV, Lattos A, Cheimonopoulou MT, Michaelidis B, Apostolidis AP, Giantsis IA (2024) First detection of *Aphanomyces astaci* and its potential responsibility for mass mortalities of *Pontastacus (Astacus) leptodactylus* in Greek lakes. *Journal of Biological Research-Thessaloniki* 31.

Anderson R, Gutreuter S (1983) Length, weight, and associated structural indices. In Nielsen L, Johnson D (eds.), *Fisheries Techniques*, 283–300. American Fisheries Society, Bethesda.

Benzer S, Benli ÇK, Benzer R (2015) The comparison of growth with length-weight relation and artificial neural networks of crayfish, *Astacus leptodactylus*, in Mogan Lake. *Jouranl of Black Sea/Mediterranean Environment* 21(2): 208–223.

Benzer S, Benzer R (2016) Evaluation of growth in pike (*Esox lucius* L., 1758) using traditional methods and artificial neural networks. *Applied Ecology and Environmental Research* 14(2): 543–554.

Benzer S, Benzer R (2017) Comparative growth models of big-scale sand smelt (*Atherina boyeri* Risso, 1810) sampled from Hirfanlı Dam Lake, Kırşehir, Ankara, Turkey.

Computational Ecology and Software 7(2): 82–90.
Benzer S, Benzer R, Günal AÇ (2017a) Artificial neural networks approach in length-weight relation of crayfish (*Astacus leptodactylus* Eschscholtz, 1823) in Eğirdir Lake, Isparta, Turkey. *Journal of Coastal Life Medicine* 5(8): 330–335.
Benzer S, Benzer R, Günal AÇ (2017b) Artificial neural networks approach in morphometric analysis of crayfish (*Astacus leptodactylus*) in Hirfanlı Dam Lake. *Biologia* 72(5): 527–535.
Benzer S, Benzer R (2018) New perspectives for predicting growth properties of crayfish (*Astacus leptodactylus* Eschscholtz, 1823) in Uluabat Lake. *Pakistan Journal of Zoology* 50(1).
Benzer S, Benzer R (2019) Alternative growth models in fisheries: Artificial Neural Networks. *Journal of Fisheries* 7(3): 719–725.
Benzer R, Benzer S (2020) Alternative approaches to traditional methods for growth parameters of fisheries industry: Artificial Neural Networks. In *Enterprise & Business Management: A Handbook for Educators, Consultants, and Practitioners*, 317.
Benzer S, Benzer R (2022) Morphometric analysis of Crayfish–traditional and artificial intelligent approach. *Thalassas: An International Journal of Marine Sciences* 38(2): 989–996.
Benzer S, Benzer R (2023a) *The Use of Artificial Intelligence Technologies in the Biological Sciences.* Contemporary Studies in Biological Sciences I. Ankara: Akademisyen Press.
Benzer S, Benzer R (2023b) Growth parameters with traditional and artificial neural networks methods of big-scale sand smelt (*Atherina boyeri* Risso, 1810). *Ege Journal of Fisheries and Aquatic Sciences* 40(2): 96–102.
Berber S, Akhan S, Bektaş Y, Kalaycı G (2020) Meat Yield and Length-Weight Relationship of Freshwater Crayfish (*Pontastacus leptodactylus* (Eschscholtz, 1823)) Population in Nine Different Inland Water Resources in Turkey. *Acta Natura et Scientia* 1(1): 82–95.
Boyalık F, Berber S, Kale S (2023) Meat Yield and the Length–Weight Relationships of the Narrow-Clawed Crayfish, *Pontastacus leptodactylus* (Eschscholtz, 1823). *Momona Ethiopian Journal of Science* 15(2): 189–215.
Crandall KA, De Grave S (2017) An updated classification of the freshwater crayfishes (Decapoda: Astacidea) of the world, with a complete species list. *Journal of Crustacean Biology* 37(5): 615–653.
Dartay M (2023) Investigation of Fishing Efficiency of Crayfish *Pontastacus leptodactylus* (Eschscholtz, 1823) in Keban Dam Lake. *Dünya Sağlık ve Tabiat Bilimleri Dergisi* 6(2): 102–110.
Dash G, Sen S, Pradhan RK, Ghosh S, Josileen J, Jayasankar J (2023) Modeling framework for establishing the power law between length and weight of fishes and a meta-analysis for validation of LWRs for six commercially important marine fishes from the northwestern Bay of Bengal. *Fisheries Research* 257: 106496.
DSİ (1968) *Limminological Survey Report of Hirfanli Dam Lake*. Ankara, Turkey (Turkish).
Ezziane Z (2006) Applications of artificial intelligence in bioinformatics: A review. *Expert Systems with Applications* 30(1): 2–10.
Froese R (2006) Cube law, condition factor and weight–length relationships: history, meta-analysis and recommendations. *Journal of Applied Ichthyology* 22(4): 241–253.
Froese R, Tsikliras AC, Stergiou KI (2011) Editorial note on weight-length relations of fishes. *Acta Ichthyologica et Piscatoria* 41(4): 261–263.

Garabaghi FH, Benzer R, Benzer S, Günal AÇ (2022) Effect of polynomial, radial basis, and Pearson VII function kernels in support vector machine algorithm for classification of crayfish. *Ecological Informatics* 72(Dec): 101911.

Gören GU, Karayücel S (2022) Determination of some population parameters of freshwater crayfish (*Astacus leptodactylus* Eschscholtz, 1823) in Tatli and Gici Lakes from Bafra fish lakes. *Turkish Journal of Agriculture - Food Science Technology* 10(5): 925–932.

Gültepe Y, Berber S, Gültepe N (2023) *Modeling and predicting meat yield and growth performance using morphological features of narrow-clawed crayfish with machine learning techniques.* bioRxiv 2023-03.

Hopfield JJ (1988) Artificial neural networks. *IEEE Circuits and Devices Magazine* 4(5): 3–10.

Köksal G (1988) *Astacus leptodactylus* in Europe. In DM Holdich, RS Lowery (eds.), *Freshwater crayfish biology, management and exploitation*, 365–400. London: Croom Helm.

Koutrakis E, Perdikaris C, Machino Y, Savvidis G, Margaris N (2007) Distribution, recent mortalities and conservation measures of crayfish in Hellenic fresh waters. *Bulletin Français de la Pêche et de la Pisciculture* 385: 25–44.

Kozák P, Füreder L, Kouba A, Reynolds J, Souty-Grosset C (2015) Current conservation strategies for European crayfish. *Knowledge Management Aquatic Ecosystems* 401: 01.

Madrigal-Bujaidar E, Alvarez Gonzales I, Lopez E, Picos R (2017) The Crayfish *Cambarellus montezumae* as a Possible Freshwater Non-conventional Biomonitor. In ML Larramendy (ed.), *Ecotoxicology and Genotoxicology: Non-traditional Aquatic Models*, 157–179.

Munro JL, Pauly D (1983) A simple method for comparing the growth of fishes and invertebrates. *Fishbyte* 1(1): 5–6.

Özcan Eİ, Serdar O (2018) Artifical neural networks as new alternative method to estimating some population parameters of Tigris Loach (*Oxynoemacheilus tigris* (Heckel, 1843)) in the Karasu River, Turkey. *Fresenius Environmental Bulletin* 27(12B/2018): 9840–9850.

Özcan EI (2019) Artificial neural networks (a new statistical approach) method in length-weight relationships of *Alburnus mossulensis* in Murat River (Palu-Elaziğ) Turkey. *Applied Ecology and Environmental Research* 17(5): 10253–10266.

Özcan EI, Serdar O (2019) Evaluation of a new computer method (ANNs) and traditional methods (LWRS and VBGF) in the calculation of some growth parameters of two cyprinid species. *Fresenius Environmental Bulletin* 28(10): 7644–7654.

Ricker WE (1973) Linear regressions in fishery research. *Journal of the Fisheries Board of Canada* 30(3): 409–434.

Roljić RL, Nikolić VP, Đikanović VĐ, Zorić KS, Urošević AM, Marković VM (2024) Morphometric characteristics of spiny-cheek crayfish *Faxonius limosus* (Rafinesque, 1817) from the Danube River on the territory of Serbia. *Archives of Biological Sciences* 76(1): 91–101.

Roul SK, Akhil AR, Retheesh TB, Rajesh KM, Ganga U, Abdussamad EM, et al. (2020) Length-weight relationships of fifty fish species from Indian waters. *Thalassas: An International Journal of Marine Sciences* 36(May): 309–314.

Samoilys MA (1997) Periodicity of spawning aggregations of coral trout *Plectropomus leopardus* (Pisces: Serranidae) on the northern Great Barrier Reef. *Marine Ecology Progress Series* 160(Dec): 149–159.

Suryanarayana I, Braibanti A, Rao RS, Ramam VA, Sudarsan D, Rao GN (2008) Neural networks in fisheries research. *Fisheries Research* 92(2–3): 115–139.

Türeli Bilen C, Kokcu P, Ibrikci T (2011) Application of Artificial Neural Networks (ANNs) for weight predictions of blue crabs (*Callinectes sapidus* Rathbun, 1896) using predictor variables. *Mediterranean Marine Science* 12(2): 439–446.

Türker D, Zengin K, Tünay ÖK (2018) Length-Weight Relationships for Nine Chondrichthyes Fish Species from Edremit Bay (North Aegean Sea). *Turk Journal of Fisheries and Aquatic Sciences* 19(1): 71–79.

Wang W, Xu Z (2004) A Heuristic Training for Support Vector Regression. *Neurocomputing* 61(Oct): 259–275.

Yazıcıoğlu B, Kouba A, Kozák P, Niksirat H (2018) Post-mating spermatophore storage strategies in two species of crayfish: implications for broodstock management. *Animal* 12(3): 554–558.

Zou J, Han Y, So SS (2009) Overview of artificial neural networks. *Methods in Molecular Biology* 458: 15–23.

Athens Journal of Sciences- Volume 11, Issue 2, June 2024 – Pages 89-108

Identifying Sustainability Efforts in Company's Reports Using Text Mining and Machine Learning

By Evangelos Xevelonakis[*] & Tanbir Mann[±]

This study delves into the utilization of text mining to scrutinize social and environmental reports of companies, showcasing its effectiveness in evaluation. It explores various text mining techniques and practically applies decision tree, k-nearest neighbors, and naïve Bayes methods. The paper offers guidance on extracting pertinent terms related to four CSR dimensions: Environment, Employee, Social responsibility, and Human rights. Results demonstrate the successful differentiation of text based on these dimensions, leveraging a CSR-relevant dictionary by Pencel and Malascue. Employing document classification techniques, the study constructs four models using distinct text mining approaches for comparative analysis. Through this research, the valuable role of text mining in assessing social and environmental disclosures is underscored, providing insights into optimizing these techniques for evaluations and emphasizing their potential to enhance understanding and decision-making in corporate social responsibility assessments.

Keywords: *sustainability, text mining, machine learning, Corporate Social Responsibility - CSR, environmental reports*

Introduction

Text mining, also known as text data mining or knowledge detection from textual sources, generally refers to the process of extracting compulsive and significant patterns or information from unstructured text documents (Tan 1999). This handling of unstructured data distinguishes text mining as a discriminative technology from newer models whose application is based more on the use of generative AI.

With text being the most common form of storing information, it is safe to say that text mining has a high potential in commercial contexts. As most of the information within companies is stored in textual form, text mining is one of the most useful means of gathering company related information. Text mining, however, is also a complex task as it includes dealing with unstructured and fuzzy text data. Text mining is a multidisciplinary field involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining (Hearst 1997).

One interesting field where text mining technology could produce important insights is social responsibility, which plays a major role in today's society. Social responsibility is not only restricted to individuals, but also to companies and most companies are expected to undertake social aspects in their business engagements.

[*]Head, Center for Data Science & Technology, HWZ Zurich University of Applied Sciences, Switzerland.

[±]Research Assistant, HWZ Zurich University of Applied Sciences, Switzerland.

This field is known as Corporate Social Responsibility (CSR). CSR is as a supervised notion whereby companies apply management notions in social and environmental concerns in their business operations and interconnections with their stakeholders. It is immensely important to have full disclosure to stakeholders, especially in issues of corporate social responsibility. Therefore, a large number of companies are getting involved in preparing sustainability reports and generally disclosing more of their impact on sustainability, specifically providing information on economic, social and environmental dimensions, which are the major measures of corporate sustainability quality – the so-called 'triple bottom line'. These reports are either voluntary or, depending on the country examined, could be a result of legislation or part of a code of practice. Such CSR reports are often based on frameworks and standards like UN Global Compact Principles, OECD Guidelines for Multinational Enterprises, GRI guidelines, ISO 26000, AA1000 and SA88000 (Aureli 2016).

Objective

The objective of this project is to use text mining methods and techniques to perform screening and selection of qualitative information from text. This project is focused on the process of document classification that helps in automating the assessment of company related data to screen for text that is specifically related to the CSR management concept with respect to its four different dimensions – Environmental responsibility, Employee responsibility, Social responsibility and Human rights.

Related Work

In recent years, an extensive amount of research has been done in this field. Joao and Paulo used Text Mining methodology (a Bayesian contextual analysis algorithm known as Correlated Topic Model, CTM) to conduct a comprehensive analysis of 246 articles published in 40 different journals between 1988 and 2013 on the subject of cause-related marketing (Guerreiro et al. 2016). Lin and Hsu (2018) conducted a study to examine the impact of corporate social responsibility (CSR) news reports on corporate operating performance forecasting using a large database of publicly listed electronics firms in Taiwan. Applying text mining techniques and latent topic modelling, they construct and measure the intensity of the CSR-corpus index (ICSRI), which can compress tremendous amounts of CSR textual information content into synthesized meaningful dimensions (Lin and Hsu 2018). In another study by Te Liew et al. (2014), text mining was used to identify sustainability trends and practices in the process industries. Four main sectors of the industry were studied: oil/petrochemicals, bulk/specialty chemicals, pharmaceuticals, and consumer products. The study reveals that the top sustainability focuses of the four sectors are very similar: health and safety, human rights, reducing GHG, conserving energy/energy efficiency, and community investment (Te Liew et al. 2014).

In an empirical study, results show that the financial report sentiment based on the PESTEL model, Porter's Five Forces model, and Value Chain (Primary and Support Activities) significantly correlates with the CSR score. The study characterizes the interaction between the sentiment analysis of financial reports and CSR scores (Song et al. 2018). Pencle and Mălăescu (2016) developed a content analytic dictionary using computer-aided text analysis (CATA). The dictionary, validated in the context of U.S. IPOs between 2011 and 2013, revealed four dimensions of corporate social responsibility. Each of these dimensions is used to predict IPO (Initial Public Offering) size in terms of offering price and total shares offered, as well as underpricing on the first day of trade. Their research provides a new measure of CSR that can be generalized to other text documents issued by a corporation (Pencle and Mălăescu 2016).

Text Mining Algorithms

Text mining algorithms are specific data mining algorithms in the domain of natural language processing (NLP). The text can be any type of content – postings on social media, email, business word documents, web content, articles, news, blog posts, and other types of unstructured data. These algorithms help to provide some understanding of how text is processed. Some text mining techniques, such as document classification and information retrieval, are concentrated on ranking or locating specific types of documents in a large document base. The main assumption behind these algorithms is that syntactic similarity (similar words) implies semantic similarity (similar meaning). However, these techniques work very well depending on the syntactic information, just because the used documents are often on the same subject and therefore share many keywords or terms (Ghosh et al. 2012).

Algorithms for text analytics incorporate a variety of techniques such as text classification, categorization, and clustering. All of them aim to uncover hidden relationships, trends, and patterns which are a solid base for business decision-making.

Document Classification

Document Classification methods are the best-known methods of Text Mining. They use training data with known target variables to classify new documents into a set of categories (Miner et al. 2012). Statistical models but also rule-based models can be used. There are two approaches, the instance-based approach (lazy learning), where new data sets are classified directly using the training data, and the model-based approach (eager learning), where a model is calculated and applied to the new data (Cleve and Lämmel 2016). These methods are used, for example, to identify spam emails or to classify customers into risk groups according to their creditworthiness.

For classification, a term-by-document matrix is used as input. The terms represent the features that should be carefully chosen, and the documents should be well described. However, other properties can also be used as features - for example, the amount of text, number of key words or the URL. Especially documents from the internet have special features that can be used for classification. For example, the URL contains information about the protocol used, the domain and the directory depth. At the end, all text-based features must be transformed into numerical features (Miner et al. 2012).

K-Nearest Neighbor

The k-nearest neighbor method is known as an instance-based method, in which new data sets are classified with the help of the training data. This approach requires a term-by-document matrix, including the target variables. Then, depending on these variables, the method tries to find the "nearest neighbor", i.e. a record in the training data that is the closest to the new record to be classified. The category of the found training data record is finally adopted. To find the nearest neighbor, similarity measures like the Euclidean distance are used. The number of "neighbors" to be used for the selection of the category must be selected in advance. At k=1, the category of the next neighbor is taken over. If k>1, the category that occurs most frequently among the neighbors used is selected (Cleve and Lämmel 2016).

Figure 1. *Addition of New Data Point*

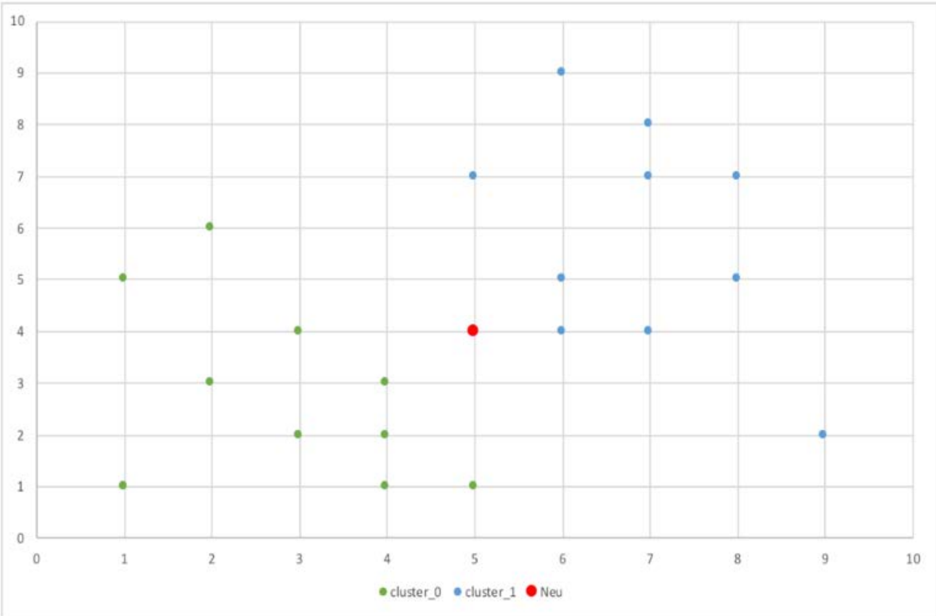


Figure 2. *Cluster Assignment*

k	x	y	Cluster
1	5	4	cluster_1
2	5	4	cluster_1
3	5	4	cluster_1
4	5	4	cluster_1
5	5	4	cluster_1
6	5	4	cluster_1
7	5	4	cluster_1
8	5	4	cluster_0
9	5	4	cluster_1
10	5	4	cluster_1
11	5	4	cluster_0
12	5	4	cluster_0

In the above example a new data point (red) is added with the coordinates (5:4) in Figure 1. The algorithm should now decide to which cluster the new point belongs - to "cluster_0" or to "cluster_1". If the algorithm is used with different values for k, the assignment to a cluster can change (see Figure 2).

Naive Bayes Classifier

The Naïve Bayes method is a probability-based method, with an instance-based approach. The goal is to predict the most probable class. The algorithm "is based on the Bayesian formula, which in calculations with conditional probabilities allows the interchanging of dependent events" (Cleve and Lämmel 2016). The algorithm assumes the independence of the variables. X is the target variable, so that $(X|Y)$ is calculated for both values and finally the value with the higher value is applied.

$$P(X|Y) = \frac{P(Y|X) * P(X)}{\mathrm{P(Y)}}$$

Decision Trees

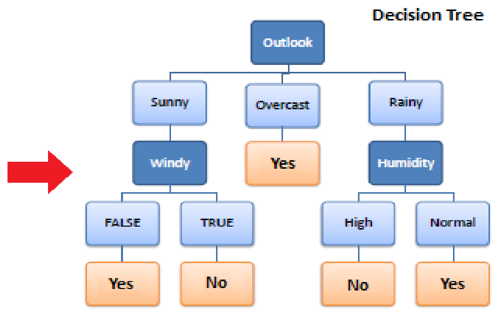
A decision tree is an instrument to display knowledge or results of a condition in a tree-like structure. The topmost node is the root, an edge represents a rule and a node which has no further branches represents a leaf. From the root to a leaf, certain rules apply, which are also applied to the data to classify it according to the leaf.

A rule imposes certain conditions (for example, greater, less or equal) on the variables of the data set, which lead to the data set being separated in such a way that one or more variables are as complete as possible in a subset. The variables that best describe the target variable are highest in the decision tree (Cleve and Lämmel 2016). The selection of the variables can be done manually, randomly or by calculation.

Figure 3. *Decision Tree Example*

Predictors / Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



In the first step, a variable is selected, and a rule is established to separate the data sets. The goal is to maximize the information content at each step. If a split is no longer possible, the branch ends in a leaf. The class of this sheet is determined by the most frequently occurring class in the sheet. The information content can be measured, for example, by the Gini index, an unequal distribution coefficient. It measures the diversity in a set and varies between 0 and 1. The lower the index, the more even the distribution.

In Figure 3, the decision tree formation is shown with a document example (Sayad 2024).

Data Collection and Preparation

In the initial phase of the project, only a single document was analyzed to check the feasibility of the text mining approach. Later, the same approach was used on around 60 documents. The data science tool KNIME was used for the implementation.

The process that has been followed in order to successfully implement the objective of this project starts with data preprocessing, where all the collected documents are preprocessed first, resulting in a clean dataset. This preprocessed data is then tagged using a CSR dictionary from Pencle and Malaescu (2016). After document tagging a bag-of-words model is used to generate the word cloud. A word cloud helps to analyze the terms that are frequently reoccurring in a document. Later, based on this bag of words creator, a document vector was generated and the terms which were present in the bag of words creator were then analyzed using 3 different text mining algorithms as described in section 3.

In the following, a step-by-step description of the analysis and the workflows that are used in the project is presented.

Data Preprocessing

The collected documents, which are in a text format also known as unstructured data, are processed and converted into a format that can be used for analysis using text mining methods. The pdf documents are cleaned in a first step to ensure that only relevant data is loaded. Subsequently, the data set is tagged by a CSR

Dictionary. First, a CSR dictionary is procured, a KNIME Extension is programmed and finally a KNIME Workflow is developed to tag the CSR relevant terms in the texts. More detailed information about the CSR dictionary is provided in the next section.

In the data cleaning process, numbers, punctuation marks and stop words are removed and all letters are converted to lower case. As a last step, all words with less than 3 letters are removed. Further pre-processing steps, such as the creation of a bag-of-words, are performed later in the application of a CSR dictionary. Stemming and lemmatization are not used as they are not required when using a dictionary.

Figure 4. *Data Preprocessing Workflow*

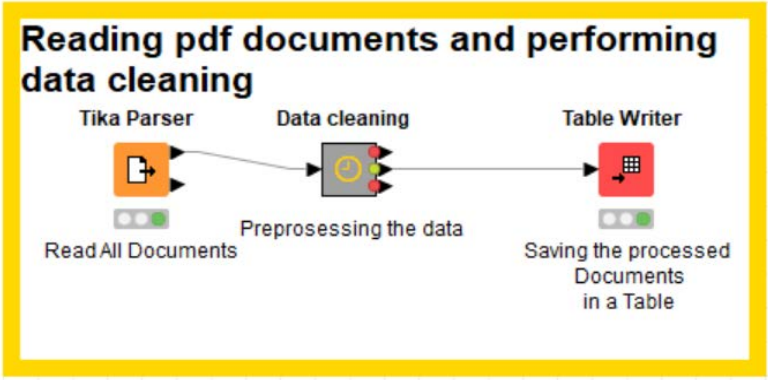
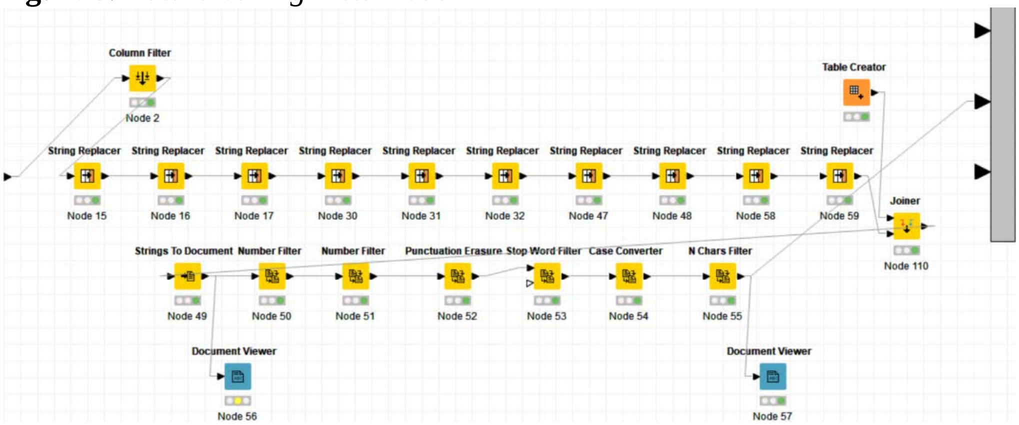


Figure 5. *Data Cleaning Meta Node*



As shown in Figure 4, a *Tika Parser* node is used to read the pdf data. In the data cleaning meta node (see Figure 5), a *String Replacer* node is used to clean the documents with patterns like: –, ®, •, etc., along with punctuations and stop words. A *Table Creator* node is utilized to categorize each document into one of four dimensions: Environment, Employee, Social, and Human Rights. All the documents are pre-classified with respect to these dimensions and hence are added in the documents pre-processing phase. This step is performed in the very beginning to get the category of each document. This step is essential as later *Document Vector* and *Document Extraction* nodes are used to create term metrics (further description is provided in the analysis section). A *Joiner* node is used to join the two tables into one and adds a new column called 'class' in the document table. At the end of the preprocessing phase, the documents are stored in a new table which is then used to perform the text mining methods.

After preprocessing, the next steps focus on tagging the terms in the clean documents with a CSR dictionary developed by Pencle and Malaescu (2016).

CSR Dictionary

Text Mining often uses lexical resources. A dictionary is a list of words that is relevant to a certain topic. It is important to use a specific dictionary for each topic. These can be simple word lists, dictionaries or even a thesaurus. The creation of such lists is very time-consuming and requires experts in the respective field. Similar to the POS-Tagger, where words are assigned to different word types, there can also be different domains/categories assigned to the words. It can happen that a word is assigned to several categories (Ignatow and Mihalcea 2018).

For CSR, Pencle and Malaescu (2016) published a dictionary that they used to predict metrics for IPOs. In doing so, they made sure that the dictionary could also be used for other analyses of company texts. The dictionary consists of four dimensions and a total of over 1,400 words. To determine the dimensions, the largest CSR reporting frameworks and guidelines were analyzed and finally the dimensions "Social and Community", "Employee", "Environment" and "Human Rights" were selected. To determine the words per dimension, an initial list was drawn up based on CSR literature, which was reviewed and supplemented by experts in several stages.

For this analysis, we used the dictionary on the given documents and counted the number of CSR-relevant words per document and per dimension. The following assumptions are made for the analysis:

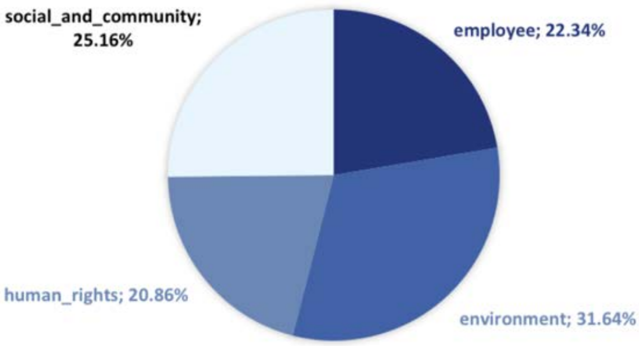
- Companies that want to go public must be attractive for investors. It can therefore be assumed that they include CSR-relevant information in their texts.
- A high CSR word count per text can be seen as an indicator of sustainable business practices and therefore leads to a higher company value (offering price).
- The companies are involved in numerous activities in various dimensions. All activities together form the CSR orientation of the company.

The dictionary is available online and can be downloaded, for example, using the R programming language. It is included in the package "lexicon" with the name "key_corporate_social_responsibility". The dictionary contains the three columns "dimension", "regex" and "token" (see Figure 6). "Token" is a placeholder for the actual word and with the help of "regex", you can easily search for the word in a text. A total of 1´421 words is included, which are distributed over the four dimensions, as shown in Figure 7. The dictionary contains tokens with only one word or tokens that are composed of several words (e.g., "water desalination"). A token can occur in several dimensions.

Figure 6. *CSR Dictionary*

S dimension	S regex	S token
environment	\bwaste reduction\b	waste reduction
environment	\bwasteland\b	wasteland
environment	\bwater\b	water
environment	\bwater desalination\b	water desalination
environment	\bwater purification\b	water purification
environment	\bwater purifications\b	water purifications
environment	\bwave\b	wave
environment	\bweather\b	weather

Figure 7. *CSR Dimension Distribution*



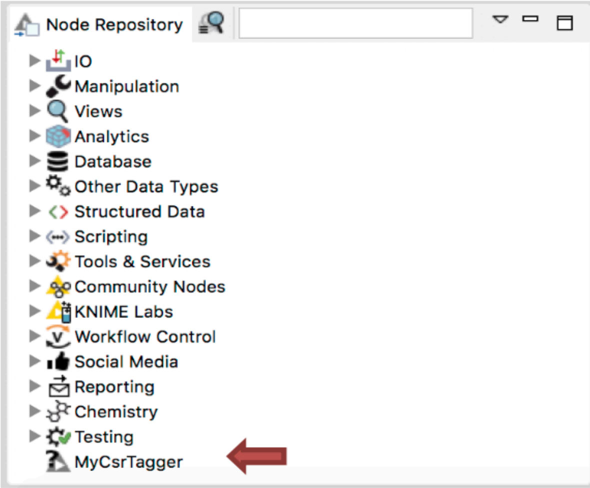
Tagging Records

In KMIME, besides the well-known POS taggers (POS Tagger, Standfort Tagger), there are also taggers for medical or chemical terms (Abner Tagger, Oscar Tagger). The *Wildcard Tagger* node can be used to tag these records and also allows you to use your own word list as input. This node has 12 tag types implemented, with the corresponding tag values to be used. These are common taggers, which are mainly used in the medical field. For CSR there is no integrated tag set available. However, KNIME offers the possibility to program your own node extension with your own tag set (KNIME 2023).

For programming the node extension, Eclipse, a development environment that is often used for Java programming, is needed. The same instructions as on the Sentimental analysis node example can be followed, only a few adjustments are required (KNIME 2023). The name of the project and the Java classes should be "MyCsrTagger", the TAG_TYPE should be "CSR" and the tagset should be the four categories from the CSR Dictionary by Pencle and Malaescu (2016): "Social_and_Community", "Employee", "Human_Rights" and "Environment".

Once all the steps in the instructions have been completed, the project must be saved as a .jar file and placed in the "dropins" folder of the KNIME installation folder. The next time KNIME is started, the new node *MyCsrTagger* will appear in the node repository (see Figure 8).

Figure 8. *MyCsrTagger Node*



The new CSR tag-set is now available in the *Wildcard Tagger* Node and in the *Tag Filter* Node, the four dimensions are provided (see Figure 9).

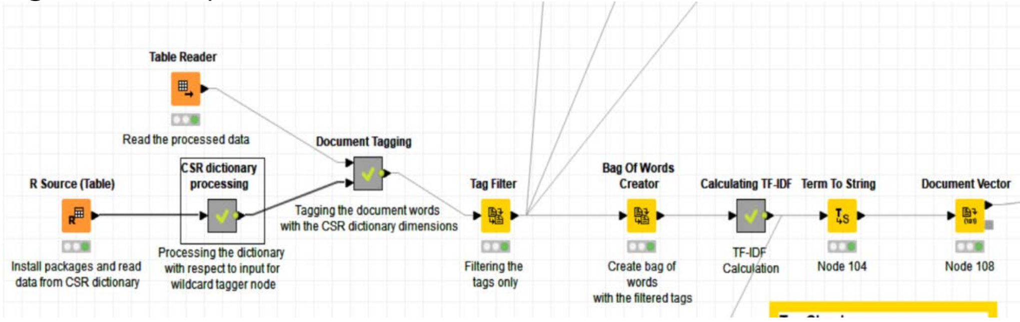
Figure 9. *CSR Tag-set*

To use the new extension with the four tag values, you have to select the category and enter the corresponding words as input into the node. But since the dictionary contains tokens with one or more words, the process must be designed accordingly. If the token with one word (e.g., "class") is found first, the token with more words (e.g., "working class"), which contains the first token, cannot be found anymore.

Therefore, the tokens in the dictionary are sorted in descending order according to the number of words per token. To implement this, the spaces per token are counted. Once these are known, the dictionary is sorted with respect to the number of spaces in descending order. This results in the tokens with several words being used to tag the documents first, while the tokens with only one word follow.

In the following workflow (see Figure 10), the dictionary is sorted into right format. The text in the preprocessed documents is tagged with the tokens in the dictionary.

Figure 10. *Workflow*



The *Document Tagging* meta-node contains the workflow where each document is tagged one by one in a recursive loop (see Figure 11). Within the Recursive Loop, the document is tagged by the *Wildcard Tagger*. Inputs are the documents and the dictionary. The *Row Splitter* passes only the top line of the dictionary to the *Wildcard Tagger*. The rest becomes input for the next iteration. The selected row is converted into variables to control the wildcard tagger and set the corresponding tag value. In each iteration only one word is tagged. The document is then returned to the start of the loop by the *Recursive Loop End* node, so that the document is tagged with all words at the end of the loop.

Figure 11. *Document Tagging Meta Node*

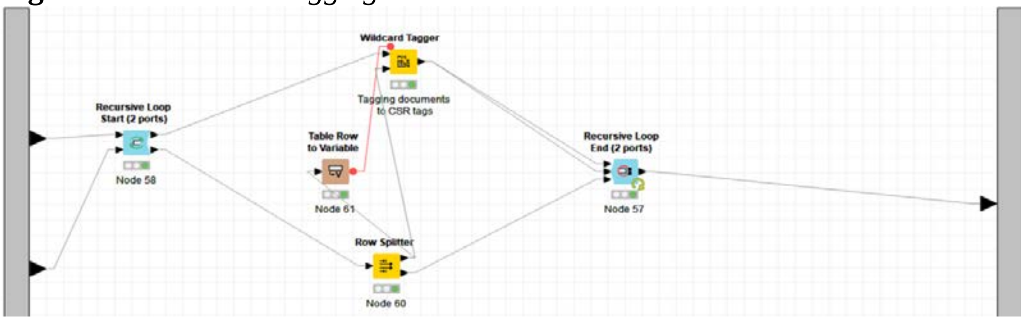


Figure 12. *Tagged Text*

UNKNOWN
responsible [CSR(Employee)] sourcing [CSR(Environment)] target [CSR(Environment)] principles [CSR(Employee)] guidelines [CSR (Environment)] design [CSR(Environment)] corporate [CSR(Environment)] design [CSR(Environment)] principles [CSR(Employee)] principles [CSR(Employee)] sourcing [CSR(Environment)] duty [CSR(Human_rights)] ownership [CSR(Human_rights)] duty [CSR (Human_rights)] duty [CSR(Human_rights)] transparency [CSR(Social_and_Community)] sourcing [CSR(Environment)] farmer [CSR (Environment)] duty [CSR(Human_rights)] duty [CSR(Human_rights)] reliability [CSR(Social_and_Community)] human [CSR (Social_and_Community)] responsible [CSR(Employee)] responsible [CSR(Employee)] employment [CSR(Human_rights)] freedom [CSR (Human_rights)] freedom [CSR(Human_rights)] employment [CSR(Human_rights)] equal [CSR(Human_rights)] respect [CSR (Social_and_Community)] safety [CSR(Human_rights)] health [CSR(Employee)] environment [CSR(Employee)] environmental [CSR (Environment)] social [CSR(Human_rights)] care [CSR(Human_rights)] nature [CSR(Social_and_Community)] stewardship [CSR (Environment)] knowledge [CSR(Employee)] transparency [CSR(Social_and_Community)] duty [CSR(Human_rights)] transparency [CSR (Social_and_Community)] duty [CSR(Human_rights)] care [CSR(Human_rights)] duty [CSR(Human_rights)] care [CSR(Human_rights)] people [CSR(Human_rights)] environment [CSR(Employee)] farm [CSR(Environment)] family [CSR(Human_rights)] farm [CSR (Environment)] responsible [CSR(Employee)] sourcing [CSR(Environment)] agricultural [CSR(Environment)] worker [CSR(Human_rights)] health [CSR(Employee)] respect [CSR(Social_and_Community)] gender [CSR(Human_rights)] empowerment [CSR(Human_rights)] principles [CSR(Employee)] seasonal [CSR(Employee)] management [CSR(Employee)] nature [CSR(Social_and_Community)] conservation [CSR (Environment)] quality [CSR(Employee)] water [CSR(Social_and_Community)] management [CSR(Employee)] practices [CSR(Employee)] farm [CSR(Environment)] water [CSR(Social_and_Community)] management [CSR(Employee)] water [CSR(Social_and_Community)] responsible [CSR(Employee)] management [CSR(Employee)] biodiversity [CSR(Environment)] management [CSR(Employee)] health [CSR (Employee)] preservation [CSR(Human_rights)] management [CSR(Employee)] farm [CSR(Environment)] cultivation [CSR (Environment)] animal [CSR(Environment)] experience [CSR(Employee)] animal [CSR(Environment)] animal [CSR(Environment)] welfare [CSR(Employee)] freedom [CSR(Human_rights)] freedom [CSR(Human_rights)] freedom [CSR(Human_rights)] freedom [CSR (Human_rights)] freedom [CSR(Human_rights)] responsible [CSR(Employee)] sourcing [CSR(Environment)] responsible [CSR(Employee)] sourcing [CSR(Environment)] responsible [CSR(Employee)] sourcing [CSR(Environment)] care [CSR(Human_rights)] respect [CSR (Social_and_Community)] enhancing [CSR(Human_rights)] communities [CSR(Human_rights)] quality [CSR(Employee)] future [CSR (Social_and_Community)] responsible [CSR(Employee)] sourcing [CSR(Environment)] requirements [CSR(Environment)] code [CSR (Environment)] responsible [CSR(Employee)] sourcing [CSR(Environment)] parties [CSR(Human_rights)] responsible [CSR(Employee)] materials [CSR(Environment)] sustainable [CSR(Social_and_Community)] agricultural [CSR(Environment)] reduce [CSR (Social_and_Community)] guidelines [CSR(Environment)] multinational [CSR(Employee)] core [CSR(Human_rights)] sourcing [CSR (Environment)] sustainable [CSR(Social_and_Community)] development [CSR(Human_rights)] goals [CSR(Employee)] services [CSR (Employee)] principles [CSR(Employee)] people [CSR(Human_rights)] communities [CSR(Human_rights)] regulations [CSR(Human_rights)] sourcing [CSR(Environment)] activities [CSR(Human_rights)] sourcing [CSR(Environment)] activities [CSR(Human_rights)] improve [CSR (Social_and_Community)] practices [CSR(Employee)] parties [CSR(Human_rights)] requirements [CSR(Environment)] help [CSR (Social_and_Community)] improve [CSR(Social_and_Community)] practices [CSR(Employee)] projects [CSR(Social_and_Community)]

Figure 12 shows a text from the analysis data set that was tagged with the CSR Dictionary. All four dimensions can be seen in the text.

Once the documents are tagged, we filter out the terms from the documents that are only relevant with our CSR dictionary and create a bag-of-words with the filtered tags. The *Bag of words creator* node helps us to collect these terms in one bag and term frequency is calculated with respect to each document to generate a Tag cloud. The Tag cloud helps to visualize the terms that have more weightage i.e., that are occurring more frequently in the documents as shown in Figure 13.

Figure 13. *Tag Cloud*

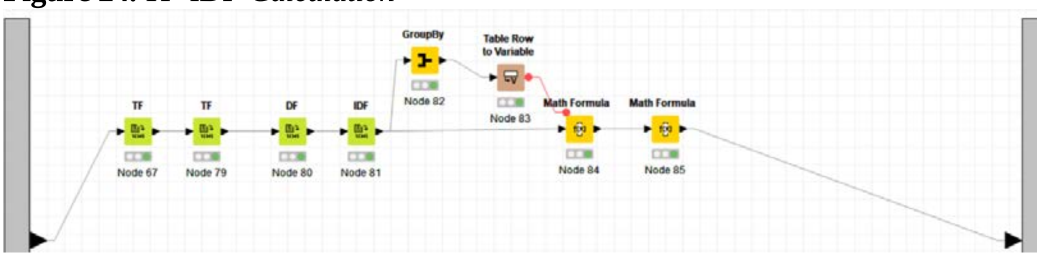
Model Development and Analysis

The models created for this analysis are based on a CSR Dictionary (lexicon-based approach). All documents have been tagged with this dictionary so that it is possible to filter according to the CSR tags. The analysis assumes that, on the one hand, only these terms are relevant for the classification and, on the other hand, that the more CSR terms occur, the more likely it is that texts are relevant.

Lexicon Based Approach with Decision Tree

With this approach, the tagged analysis data is first used to calculate the key figure TF-IDF per term. For the calculation of the TF-IDF ratios per term, the ratios TF_abs, TF_rel, DF_abs, DF_rel and IDF are calculated in the *Calculating TF-IDF* metanode (see Figure 14). To reduce the features, i.e. the terms, two filters are used. The first filter reduces the features using the key figure TF_abs. As second filter, the DF_rel value per term is used. Only terms with a DF_rel value between 0.1 and 1.0 are retained.

Figure 24. *TF-IDF Calculation*



Finally, as shown in Figure 15, the data is converted into a document matrix using the *Document vector* node. *Document data extractor* is used to extract the class of the documents which has been added manually during the data preparation phase for classification of the documents. Finally, the analysis dataset is divided into two parts by the node *Partitioning*. In this case, 70% of the data is used for the learner and 30% for the predictor. For this purpose, stratified sampling, i.e. that the distribution of the target variable, is the same in both parts, and a random seed (1578008215742) is chosen so that the results are repeatable.

Figure 15. *Decision Tree Workflow*

The Decision Tree Learner creates a model which is used for the training data to measure the quality and for the test data to make a prediction. In the node configuration, 'Gain ratio' is provided as a quality measure instead of 'Gini index' for better performance. The model has an accuracy of 0.529 and a Cohen's Kappa value of 0.352. These values are acceptable and therefore, the model can be used for making predictions. Figure 16 shows the decision tree for the model. In total, there are only seven leaves, which are created by dividing the documents by the words "employment", "cultures", "funds", "environmental", "award", "goal" and "humanitarian".

For the above analysis, it was described how a predictive model was built to predict the labels of documents with respect to CSR tags (environment, employee, social and human rights). In this approach, single words were used as features. It is shown that the most discriminative terms w.r.t separating the four classes are "environment", "waste", and "rights". If the term "waste" occurs in a document, it is likely to belong to the environment class. If "goal" occurs, it is equally likely to belong to the environment or employee class, etc.

Figure 16. *Decision Tree*

N-gram Approach

The single word approach may have several drawbacks. It is very likely that a single term analysis could lead to misclassification of data (Thiel 2016). Therefore, using frequent n-grams as features in addition to single words can overcome this problem. In order to analyze the terms that provide better understanding of the documents, the n-gram approach is used.

In this section, 1- and 2-grams as features for prediction are used in order to check if a 2-gram approach helps to increase the accuracy of the model in comparison with only single word features (see Figure 17). The KNIME Text Processing extension is used in combination with the traditional KNIME learner and predictor nodes to process the textual data as well as build and score the predictive model. For the analysis, the same approach is used as in the previous model for data preprocessing and after tagging the data, a *N-Gram creator* node is used to generate the terms that are more likely to occur together than separately (see Figure 18). A slight change is made to the configuration of the decision tree predictor node, where the 'Gini index' is used instead of the 'Gain ratio' in 2-gram and 1- and 2-gram approaches in order to achieve better accuracy of the models.

Figure 17. *N-grams*

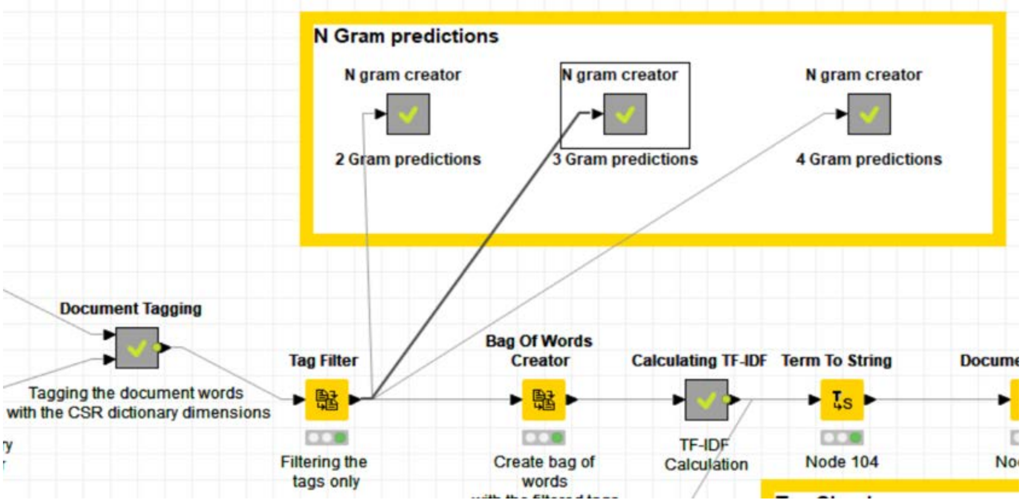
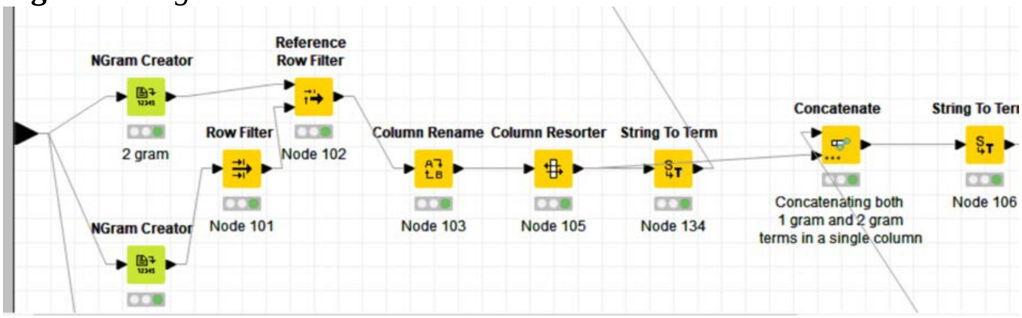
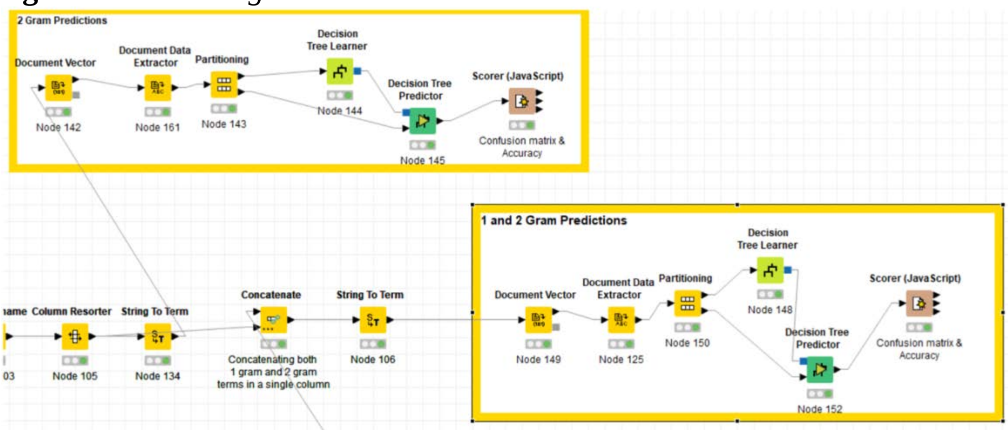


Figure 18. *N-gram Creator*



In the 2-gram predictions, the meta node has two input ports and the *String to Term* node has two output ports. In the top output port, document vectors are provided with two-word features only. In the bottom output port, document vectors with 1- and 2-gram features are provided (see Figure 19). These vectors are used for classification in the next steps. The vectors containing 2-grams only, consist of 265 features. The vectors with 1- and 2-grams consist of 551 features. Taking the 2-grams into account doubles the feature space in this case.

Figure 19. *1- and 2-gram Predictions*



For both branches (top with 2-grams as features only, bottom with 1- and 2-gram features), predictive models are built, tested and compared. For classification, any of the traditional mining algorithms available in KNIME can be used (e.g. decision trees in this case). When the documents have been created, the target variable is stored as a category in the documents. To extract the target variable as a column, the *document extraction* node is used. The *Partitioning* node splits the dataset into training (70%) and test (30%) sets in both branches.

At the end, two decision trees are built on the training sets and applied on the test sets (see Figures 20-21). The decision tree in the top branch, trained on the data set with 2-gram features, only achieved an accuracy of 0.529. The decision tree of the bottom branch, trained on the data set with 1- and 2-gram features, achieved an accuracy of 0.588. Cohen's kappa delivers a value of 0.346 and 0.449 respectively.

Figure 20. *1- and 2-gram Decision Tree*

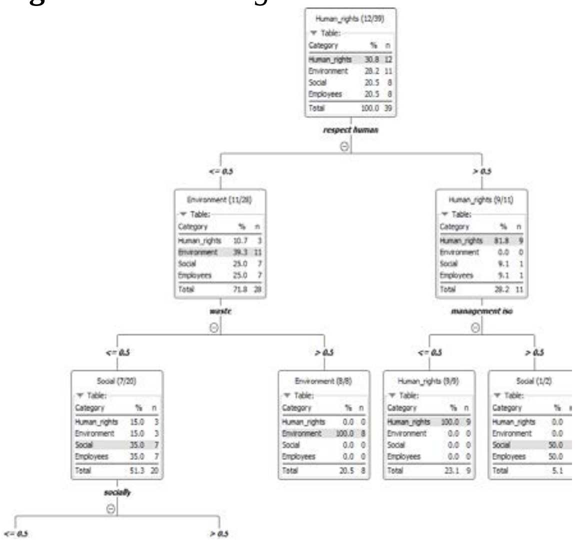
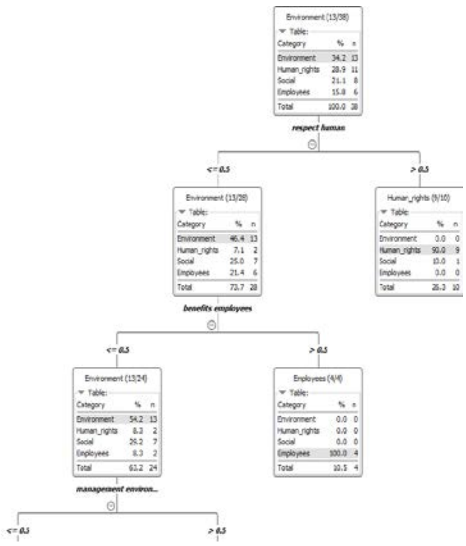


Figure 21. *2-gram Decision Tree*

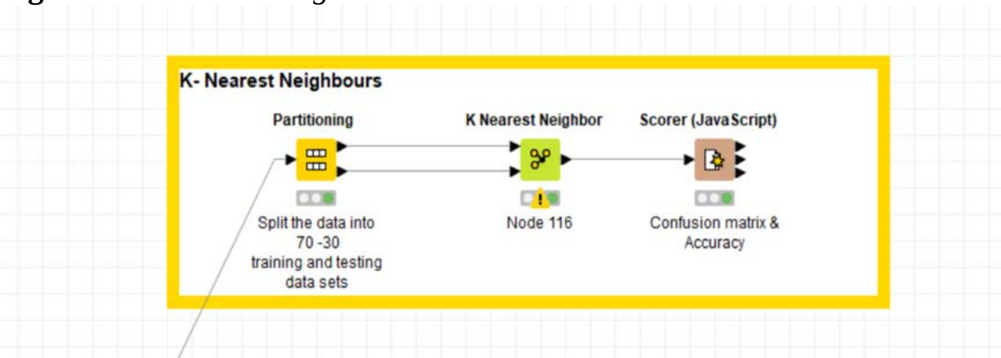


The same approach is applied on 3- and 4- gram models which resulted in less accuracy and statistically insignificant results.

Lexicon Based Approach with K-nearest Neighbor

As in the previous approach, the data set for this approach is converted into a document matrix, with the TF-IDF values as elements of the matrix. Again, the columns can be harmonized. To test the model, the analysis data is separated by the node *Partitioning* (see Figure 22). For this approach, 70% of the data is used for training the model and 30% for testing. For the k-nearest-neighbor approach the 4 nearest neighbors (k=4) are used for prediction. Although this number is not odd, as generally odd numbers are preferred for this approach, it provides the best results nonetheless (a test with different values for k was performed). With these settings, an accuracy of 0.471 and a Cohen's Kappa of 0.292 could be achieved. These results indicate that this approach is acceptable for the classification of training agendas but provides less statistically significant results in terms of Cohen's kappa and accuracy.

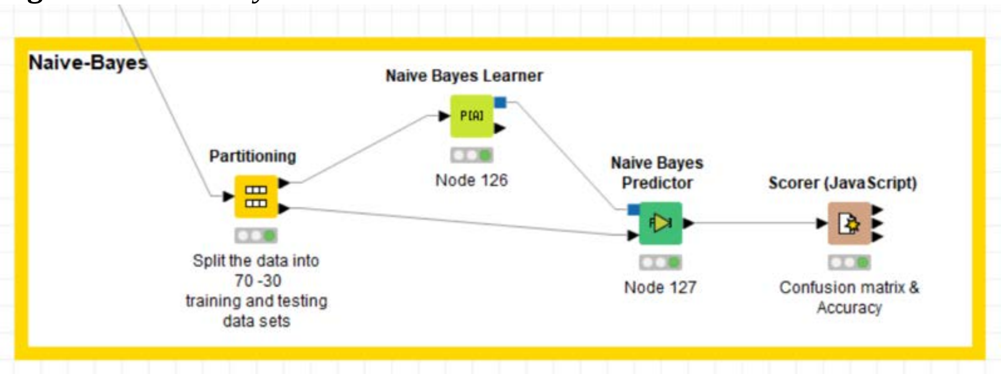
Figure 22. *K-nearest Neighbors*



Lexicon Based Approach with Naïve Bayes

As in the two previous approaches, the data sets are brought into a document matrix and the columns are harmonized. For this approach, the data is again divided into 70% for training and 30% for testing by the *Partitioning* node. Again, the stratified sampling approach is used, with a random seed (1610365614571). The Naive Bayes Learner creates a model that is used in the Naive Bayes Predictor (see Figure 23). For the training data an accuracy of 0.527 and a Cohen's Kappa of 0.355 can be achieved. These results show that the model does achieve good classifications in comparison to the two previous models.

Figure 23. *Naive Bayes*



Comparison of the Models

A comparison of the results of these different models shows that except for the K-nearest neighbor model, the values for Accuracy and Cohen's Kappa are comparable (see Table 1). The decision tree model with a 1-and 2- gram approach delivers the best results.

Table 1. *Model Comparison*

Model	Accuracy	Cohen's Kappa
Decision Tree	0.53	0.36
2-gram (decision tree)	0.53	0.35
1 & 2 gram (decision tree)	0.59	0.45
K-nearest neighbors	0.47	0.30
Naïve Bayes	0.53	0.36

Conclusion and Future Work

The importance of Corporate Social Responsibility has increased over the last years. Nowadays, large, established companies often include CSR in their strategy due to the ever-increasing pressure from the public. The models developed in this paper help us to identify the CSR relevant terms in a company's report.

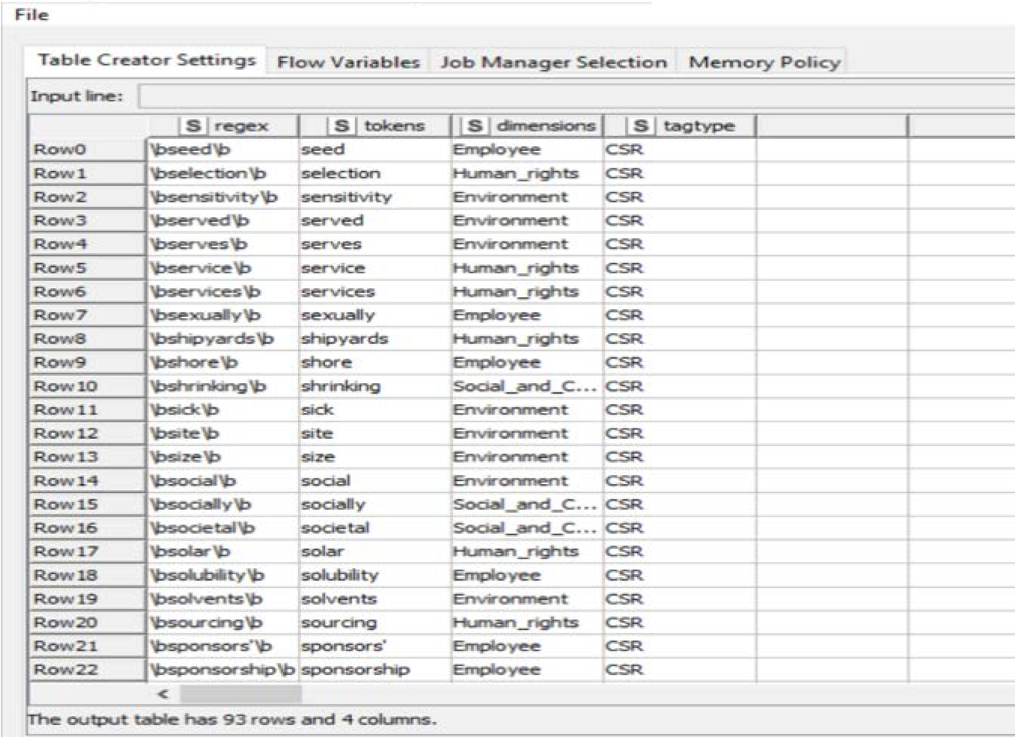
The basis of the analysis is a CSR dictionary published in 2016 by Pencle and Malaescu (2016). We used this dictionary to analyze documents for an IPO. In a first step, the CSR terms in the dictionary were divided into four dimensions, which are later used as target categories for the analysis. The use of the mentioned dictionary resulted in a reduction of number of words per text, improved workflow performance and general improvement of models based solely on these words.

Three of the four used models (N-gram models, Decision tree and Naive Bayes) showed good results. These models had an accuracy of around 53% – 59% and Cohen's Kappa values of over 0.45 when applied to the analyzed data. Only the model with k-nearest-neighbor was less accurate. Data classification by this model should therefore be interpreted with caution. The decision tree with 1-gram models showed more descriptive results with more branches and leaves with equal weightage on both the left- and the right-hand side of the tree. Whereas, in 2-grams only and 1- and 2- grams models, the decision tree is less descriptive and has more weightage on left hand side of the trees.

In order to validate the aforementioned results and to know why the trees are so different from when using the 1 term or 1-gram approach, a more in-depth analysis is needed. Furthermore, it is important to bear in mind that the success of a model depends heavily on the quality and quantity of the pre-classified analysis data. To obtain more reliable results, it is recommended that to use a larger dataset. This is also shown by the fact that the choice of a different random seed in the partitioning of the analysis data leads to very different results (for example when using the k-nearest-neighbor model).

For future work, large data sets are recommended for the validation of the above models. It is also possible to enhance the CSR dictionary with more terms which can map to different dimensions as per the business case. For enhancing the terms in the dictionary, a table must be created with the exact columns as in the CSR dictionary (see Figure 24).

Figure 24. *Creation of CSR Dictionary with New Terms*



File

Table Creator Settings | Flow Variables | Job Manager Selection | Memory Policy

Input line:

	S regex	S tokens	S dimensions	S tagtype
Row0	\bseed\b	seed	Employee	CSR
Row1	\bselection\b	selection	Human_rights	CSR
Row2	\bsensitivity\b	sensitivity	Environment	CSR
Row3	\bserved\b	served	Environment	CSR
Row4	\bserves\b	serves	Environment	CSR
Row5	\bservice\b	service	Human_rights	CSR
Row6	\bservices\b	services	Human_rights	CSR
Row7	\bsexually\b	sexually	Employee	CSR
Row8	\bshipyards\b	shipyards	Human_rights	CSR
Row9	\bshore\b	shore	Employee	CSR
Row10	\bshrinking\b	shrinking	Social_and_C...	CSR
Row11	\bsick\b	sick	Environment	CSR
Row12	\bsite\b	site	Environment	CSR
Row13	\bsize\b	size	Environment	CSR
Row14	\bsocial\b	social	Environment	CSR
Row15	\bsocially\b	socially	Social_and_C...	CSR
Row16	\bsocietal\b	societal	Social_and_C...	CSR
Row17	\bsolar\b	solar	Human_rights	CSR
Row18	\bsolubility\b	solubility	Employee	CSR
Row19	\bsolvents\b	solvents	Environment	CSR
Row20	\bsourcing\b	sourcing	Human_rights	CSR
Row21	\bsponsors'\b	sponsors'	Employee	CSR
Row22	\bsponsorship\b	sponsorship	Employee	CSR

The output table has 93 rows and 4 columns.

In order to customize the tagging dimensions, one can e.g. build any other dimensions instead of ‘environment’, ‘employee’, ‘human rights’ and ‘social responsibility’. This can be achieved by simply building a node extension and following the same steps as described in the section on CSR Dictionary. It is recommended to pay more attention during the data cleaning phase as occasionally, identical terms (e.g., ‘time time’ or ‘human human’) will occur in the decision tree while using the 2-gram approach. This could lead to confusion and doubts in the descriptiveness of the models. So, for future work, it is recommended to pay attention on data when applying a 2-gram approach.

Finally, it is safe to say that this work serves as a strong base for future projects that are focused on analyzing text data for different CSR dimensions and gathering qualitative information with respect to these dimensions.

References

Aureli S (2016) A comparison of content analysis usage and text mining in CSR corporate disclosure. *The International Journal of Digital Accounting Research* 17: 1–32.

Cleve J, Lämmel U (2016) *Data mining* (2nd ed.). De Gruyter Oldenbourg.

Ghosh S, Roy S, Bandyopadhyay SK (2012) A tutorial review on Text Mining Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering* 1(4): 223–233.

Guerreiro J, Rita P, Trigueiros D (2016) A text mining-based review of cause-related marketing literature. *Journal of Business Ethics* 139(1): 111–128.

Hearst MA (1997, July) *Text data mining: Issues, techniques, and the relationship to information access.* Presentation for UW/MS workshop on data mining, Berkeley, CA, United States.

Ignatow G, Mihalcea R (2018) *An Introduction to text mining: Research Design, Data Collection, and Analysis*. Los Angeles, CA: Sage.

KNIME (2023, November 22) *Create a New Java based KNIME Extension*. Available at: https://docs.knime.com/latest/analytics_platform_new_node_quickstart_guide/index.html#_introduction.

Lin S-J, Hsu M-F (2018) Decision making by extracting soft information from CSR news report. *Technological and Economic Development of Economy* 24(4): 1344–1361.

Miner G, Delen D, Elder J, Fast A, Hill T, Nisbet R (2012) *Practical text mining and statistical analysis for non-structured text data applications*. 1st Edition. Elsevier.

Pencle N, Mălăescu I (2016) What's in the words? Development and validation of a multidimensional dictionary for CSR and application using prospectuses. *Journal of Emerging Technologies in Accounting* 13(2): 109–127.

Sayad S (2024) *Decision Tree – Classification*. Available at: https://www.saedsayad.com/decision_tree.htm.

Song Y, Wang H, Zhu M (2018) Sustainable strategy for corporate governance based on the sentiment analysis of financial reports with CSR. *Financial Innovation* 4(2).

Tan AH (1999) Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge discovery from advanced databases*. Beijing, 8, 65–70.

Te Liew W, Adhitya A, Srinivasan R (2014) Sustainability trends in the process industries: A text mining-based analysis. *Computers in Industry* 65(3): 393–400.

Thiel K (2016, January 8) *Sentiment Analysis with N-grams*. KNIME. Available at: https://www.knime.com/blog/sentiment-analysis-with-n-grams.

Athens Journal of Sciences- Volume 11, Issue 2, June 2024 – Pages 109-124

Generative Urban Design in the Field of Infrastructure: An Optimizing Solution for Connecting Fier and Vlora County by a 600 m Bridge over Selenica River, Albania

By Ilda Rusi[*] & Albi Alliaj[±]

The way we think about infrastructure is being completely changed by parametric and generative design. Meanwhile the contemporary urban planning process is often viewed as a complicated and fragmented workflow. The main goal is to optimize solutions with tens of thousands of variations while concurrently taking into consideration various limits. This Paper will discuss and demonstrate the use of a generative urban design framework at the local scale. Although relevant to infrastructure, generative design is not limited to architecture. And on the other hand, the construction sector is becoming more specialized and complex. The close cooperation between structural engineers, architects, urban planners and other stakeholders is a major driving force behind modern projects. The building site is cut off from architects and engineers, particularly in the digital age. To ensure a three-dimensional scope of work, digital models are therefore necessary. The difficulty is that the structural model and architectural model do not match up exactly. A generative design is therefore explained in the case of a bridge design. Bridges are effective structures that provide a variety of topologies, materials, and geometries. This paper examines how the geometry and topology of a 600 meters long bridge can bring an optimal solution for connecting two nearby counties, Fieri and Vlora. The performance of the bridge can be examined by altering the geometrical parameters in addition to the topology. By adding more design factors and offering a fresh method for bridge optimization, the study aims in further developing the initial parametric model. Since the process of changing the design is quite quick and the analysis is displayed instantly, using parametric design to study alternative options for bridges could be highly helpful to designers.

Keywords: *generative design, bridge, geometry, typology, optimization, deflection.*

Introduction

The generative urban works in fact count for many urban aspects and corresponding structures. One of them is related to terrain modelling and road planning which several times go hand in hand. However, developing them in 2D poses numerous difficulties and frequently results in time lost through rework. Road and infrastructure engineers may need to start the same project over more than once due to their conflicting perspectives and that of numerous other stakeholders. This re-work can be avoided by using a parametric road modelling

[*]Deputy Dean, Faculty of Architecture and Design, POLIS University, Albania.
[±]PhD Candidate, POLIS University, Albania & University of Ferrara, Italy.

software solution, which makes road design and terrain modelling considerably simpler and more effective. The road body can be modelled once the terrain's surface and axes have been established. There are several plugins that help towards employing a template-based modelling strategy, which has the benefit of being highly flexible.

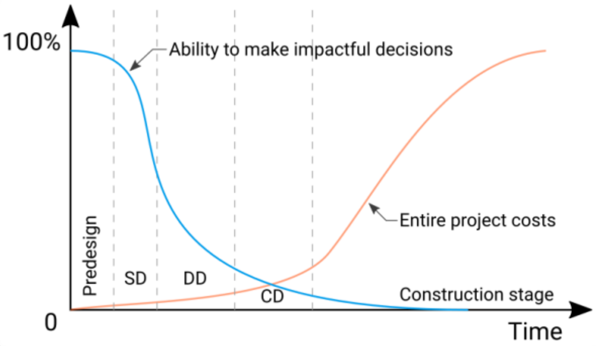
There are four templates that define the fundamental road geometry:

- The road layer structure, standard width, vertical offsets, and lane count are all specified in the cross-section template.
- All of the parameters for the substructure are defined in the sub-base template.
- Every shoulder, ditch, slope, and wall are defined using the roadside templates.
- The calculation parameters for the model are defined by the land requirement template.

The templates are allocated to the axis after they have been constructed. Of course, by setting the inlet and outlet distances, the transitions between adjacent templates are automatically made. A road body's geometry can vary a lot and in this regard the plugins offer numerous detailing functions to improve the structure, preventing the need for hundreds of templates for tiny variations. The templates allow for the overwriting of all previously set parameters to meet specific criteria. The above example of the road preliminary design aspect are very much related also with the bridge design as one important element of the whole terrain modeling.

Nevertheless, it is important to intiate explaining the broad concept of the generative urban design. And in this regard, the three main elements of the generative design process are based on how natural processes create shapes in complicated patterns (Krish 2011). There are several advantages to using these essential elements in product design. The first step is to create design solutions that are effective, resilient and compatible. The use of digital technologies and algorithms to generate a huge number of workable design directions comes in second. The third step is to design beautiful, dynamic forms and patterns.

Figure 3. *Predesign and Schematic Design Stages are when the Majority of Critical Decisions that Affect the Design and Cost of Construction are Made*



Source: BLOG on AEC Innovation. https://www.invokeshift.com/thoughts-on-the-future-of-generative-design-in-aec-from-an-engineering-perspective/.

There are numerous ways to integrate generative design into the design process and by the perspective of a designer, the majority of methods fall into two broad categories: *i) by subtraction* and *ii) by addition*. Parts of a product are examined based on their strength or durability during a subtractive process in order to remove superfluous pieces while preserving performance. Utilizing techniques like shape optimization, trabecular structures, and lattice design, this strategy is accomplished (Autodesk, Inc 2018, Singh and Gu 2012). A subtractive technique has a very short learning curve, although it also originates from too precise designs and offers only minor advancements over existing options. An additive technique produces a wide number of viable solutions that satisfy the design objectives and limitations for a particular challenge.

There are many techniques for additive generative design, including tabernacle structures. Although the generative design far exceeds what human capability could produce on its own, it also makes it difficult to develop algorithms that function as expected, let alone to choose viable and desirable solutions from among the many options generated (Cui and Tang 2012). The role of the designer, who works with computers to create systems that are sophisticated, linked, robust, and novel, is at the centre of this process (McCormack et al. 2004). The authors in fact mention also the iterative design process, a very interesting approach to the design phase too. The relationship between generative design and iterative design process offers an innovative framework where designers may feel at ease utilizing tried-and-true workflows and fusing them with cutting-edge technology, all of which results in more effective and compelling designs.

New strategies and technologies are required to help urban designers plan resilient and sustainable urban landscapes. Numerous computational methods have been suggested, such as automated production of urban design suggestions based on predetermined parameters or various types of spatial analysis to assess the effectiveness of design plans. However, the majority of these ideas have led to isolated tools and disjointed workflows. An appropriate computational representation of the urban design problem is one of the primary obstacles to merging urban analytics and generative approaches in the framework of urban design optimization procedures.

A comprehensive data representation for urban fabrics, including the organization of street networks and parcels, is offered to help overcome this challenge. This form may be effectively employed with evolutionary optimization techniques. It is shown how the data structure developed for the Grasshopper for Rhino3D software can be used as a component of an adaptable, modular, and extensible optimization system that can be applied to a range of urban design issues and can reconcile potentially incompatible design objectives in a semi-automated design process. The proposed case study method intends to help a designer by introducing possibilities into the design phase for deeper investigation. An urban design concept for the communities of Fieri and Vlora is used to illustrate how the system works.

The paper's objectives are to automate bridge modeling in the early design phases, to conduct a full analysis, and to optimize the bridge structure with regard to reinforced concrete material used. According to the study case's findings, designing bridges utilizing optimization and parametric design has some promise. Generative designs can be investigated with minimum effort from the designer with a strong parametric design description. A parametric design approach might be used for its ability to expedite the design process and give the designer access to adaptive design.

In order to analyze the effects of the bridge link for the counties of Fieri and Vlora over the Selenica River, a generative design approach was suggested in this regard. In addition to providing deeper understanding of potential conflicts and trade-offs between design goals, the case illustrates how the Generative Design method can produce effective design strategies. It was possible to demonstrate how, at the heart of generative design, there is always a choice of inputs and constraints that do not absolve the engineer or architect of responsibility but that can be gradually refined by further enhancing the generated strategies and resulting in a more informed design. This can be done by expanding the process to include the generative design for architectural space planning's evaluative component and outlining a new set of metrics for the automatic evaluation of end-user satisfaction inside the defined areas. As a result, when big datasets are available, machine learning can be a useful tool to enhance generative design. In terms of technology, it can attempt to use machine learning at any point during the generative design process.

The Historic Path

Study the past if you want to define the future, advised Confucius. We could get a clear picture of the development of design optimization techniques in architectural and urban design from a summary of historical development (Table 1). From this image, it is anticipated that the benefits (Table 2) and difficulties (Table 3) of the methodologies will be easier to comprehend from a historical standpoint. The development of design optimization methods and strategies to deliver the newest cutting-edge technology to satisfy the continuously changing requirements in architecture and urban design is equally fascinating to observe. This part should respond to the first query.

Prior to being scaled up to urban design, design optimization first appeared in architectural design. Attempts to use optimization techniques to solve design problems may be traced back to 1969 in academia thanks to Simon's ground-breaking paper on the "Science of Design" in his influential book "The Sciences of the Artificial" (Simon 2019). The optimization process was encouraged to be one of the many attempts to demonstrate the scientificity of architecture in addition to its inherently aesthetic aspect during this time, when architecture did not even have a well-established theory (Widdowson 1971). Architects navigate through and add elements one at a time to a rich combinatorial space, which Simon (1975) further characterized as the essence of design creativity. This definition perfectly aligns with mathematical optimization.

Table 1. *The Development of the Methods used and Respective Instruments into Design Process*

Objects	Developments
1960s-1980s	the objective of a single optimization
1990s	methods instruments based in human logic
2000s	multi-objective simulation-based methods
2010s	methods using artificial intelligence

Source: Miao et al. 2020.

Table 2. *The Advantages of the Optimisation Methods in Different Periods of Time*

Objects	Advantages
1960s-1980s	the scientific framework of architecture in the architectural design
1990s	field design approach and new application methods in regard
2000s	emerged problems which are complex and the corresponding applications in urban design
2010s	technologies used in artificial intelligence and best practices to design phase

Source: Miao et al. 2020.

Table 3. *The Methods of Optimisation and the Challenges in Different Periods of Time*

Objects	Challenges
1960s-1980s	lack of mathematical models
1990s	CAAD community debates on topic
2000s	thriving scenarios
2010s	misalignment of design techniques and data-driven approaches

Source: Miao et al. 2020.

A variety of optimization techniques' uses in architecture and urban planning were explored by Gero (1975). He emphasized how the lack of numerical models in architecture limited the use of this method in design. For decades, efforts to create optimization-based design methodologies in the CAAD field persisted. Through a number of articles in the 1980s, the optimization in the design was first introduced by Gero and Radford (1984), Radford abd Gero (1987), and Balachandran and Gero (1987) to the design field. During this stage, design optimization did succeed in resolving a few related architectural design issues. However, when the design difficulties could not be expressed mathematically, the applied numerical optimization methods frequently failed.

More logic-based AI approaches were created in the 1990s to loosen the restrictions of scientific formulation, but the need for such techniques in design was hotly contested at the time. An intelligent computer-aided design system prototype that placed an emphasis on the collaboration between a computer and a person was proposed by Pohl et al. et al. (1990) and Schmitt and Oechslin (1992) drew attention to the fact that the research frontiers of CAAD were beginning to switching to design support from design automation, urged additional insights into human cognition. A design-oriented approach was put up as a way to assess, criticize, and optimize building energy use and design. Despite numerous attempts during this decade, the CAAD field has faced skepticism and criticisms.

It was indisputable that the created systems' applicability was still very constrained. Internal criticism from the field was also expressed at the same time, which pushed the research agenda forward. The CAAD's seven deadly sins, which Maver (1995) mentioned and which include macro-myopia, d'ej'a vu, xenophilia, unsustainability, failure to validate, failure to evaluate, and failure to criticize, are a well-known example. New approaches, like using genetic programming to explore design spaces, started to emerge as a result to both internal and external criticisms (Broughton et al. 1997). As processing power increased at the start of the twenty-first century, additional derivative-free and stochastic optimization techniques were developed and used to tackle challenging discrete nonlinear issues.

Utilizing generative algorithms to create architectural design shapes is one of the projects Coates et al. (2001) started in Center for Environment and Computing in Architecture. Derix (2009) employed Quantum Annealing to determine desired adjacencies between various land use units and Ant Colony Optimization to create roadway networks for urban design. In addition, fresh perspectives on the application of optimization techniques arose. Using optimization, Bleiberg and Shaviv (2007) improved collaborative design. Multi-objective optimization is also used, research also made strides, leading to the development of outstanding algorithms like SPEA2, NSGA-II, and later HypE (Zitzler et al. 2001, Deb et al. 2002, Bader and Zitzler 2011). Although more and more physical realities, like the Science City Zurich, were made possible with the aid of CAAD, design support tools did not make a substantial impact on design practice (Schmitt 2004).

Within the CAAD discipline, research in design optimization has advanced over the last ten years, moving from architectural design to urban design. Numerous studies in this field centered on exploring the design space. Turrin et al. (2001) created a technique for design exploration a combination of performance-driven geometries parametric modeling and genetic algorithms. Additionally, Stouffs and Rafiq (2015) put forth strategies for fusing generative and evolutionary exploration. When the issue is or can be reformulated as a single objective optimization problem, model-based optimization has been shown to be a faster and more practical alternative to evolutionary algorithms. Hybrid approaches using both metaheuristic and model-based optimization would suitable for multi-objective optimization issues, as has already been demonstrated in other engineering design domains (Sindhya et al. 2012).

A growing variety of quantitative evaluation methods have been brought to urban design with the advent of spatial analysis tools like space syntax, which expand the design requirements that design optimization could meet (Hillier 2007). For land use planning, Cao et al. (2011) employed multi-objective optimization techniques. In contrast to earlier design optimization attempts, their methods reinforce how designers may highlight the value of human intelligence interactions with the generated urban design. This aims to address a major criticism of computational creativity, namely the lack of humanity (Colton et al. 2014). The EMO-based generative methodologies at the heart of their strategy have the potential to improve urban design through benefits like transparency and integrativeness (Monizza et al. 2012, Singh and Gu 2012).

Additionally, this study intends to overcome the representation problem, one of the main obstacles to generative urban design. In their most recent study, the application of EMO to the creation of numerous urban design layouts including urban elements such roadway networks, blocks, lots, and buildings was successful (Koening et al. 2020). Despite progress across academic frontiers, there is still a long way to go in terms of implementation. The absence of quantitative design evaluation metrics and measures continues to be a major problem. Additionally, computer-generated design solutions are frequently straightforward and only appropriate for prototype. Because of the nature of EMO, complex processing is frequently needed even for basic generation, which takes longer than real time. A hybrid strategy is anticipated to be used to resolve these issues.

Methodology

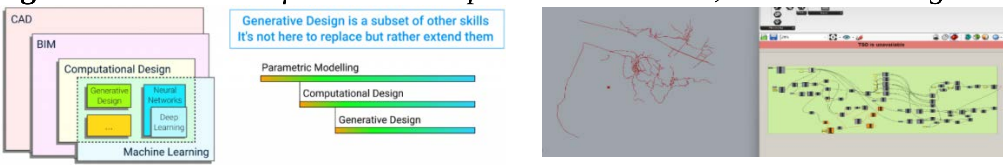
The methodology employed in this work is associated with parametric design, which is defined as a technique for producing geometries based on various parameters and rules in an algorithmic manner. The method will generate a new version of the geometry when the parameter values are altered. With the Rhinoceros add-on visual programming tool grasshopper, parametric design may be utilized. The Rhinoceros viewport is used to preview the geometry once it has been defined in Grasshopper. The structure of the model is one advantage of parametric design. A modification to the parameter will have an impact on the remainder of the design and the model if a collection of parameters defines the curves that serve as its foundation. In this regard, the chosen case study is being further analysed by primary choosing the structure, that of a 600 meters long bridge. After that the technical parameters of the bridge are being selected and the referenced schematic view of the 3D bridge is conceived. The structural sustainability is checked for the modelled bridge by resulting at the end in the proposed version for the connection of the two ground areas divided by the river.

How Does Generative Design Work?

In order to create the normal design process, also known as the traditional design process, calls upon the expertise of the designer. This is frequently a time-consuming process that necessitates designers and engineers to thoroughly comprehend many ideas and processes in order to produce an effective final phase. And then, after spending countless hours on design and analysis, the need to shorten the process frequently leads to less-than-ideal designs. Here is where generative design will play a role in creating the future's most optimal designs. This can be added to the new generation of products that are being developed, which have features of ultra-high performance and are too demanding for the conventional design process.

Designers and engineers may now co-create ideas utilizing parameter driven optimization because to the development of technology like artificial intelligence algorithms and limitless computing, which are far more accessible than at any other time in history. To assist create an optimum solution that satisfies the design goals within the constraints specified in the study setting, generative design tests the structure with each iteration, learns from each step, and applies change at each level. This technique frequently yields designs that the conventional design process would not have been able to produce. The final forms' shapes are distinctive and are referred to as "organic" because they are created to meet a particular requirement.

Figure 4. *As a Subset of Various Disciplines and Skill Sets, Generative Design*



Source: (left scheme) BLOG on AEC Innovation. https://www.invokeshift.com/thoughts-on-the-future-of-generative-design-in-aec-from-an-engineering-perspective/, (right scheme) the authors, 2023.

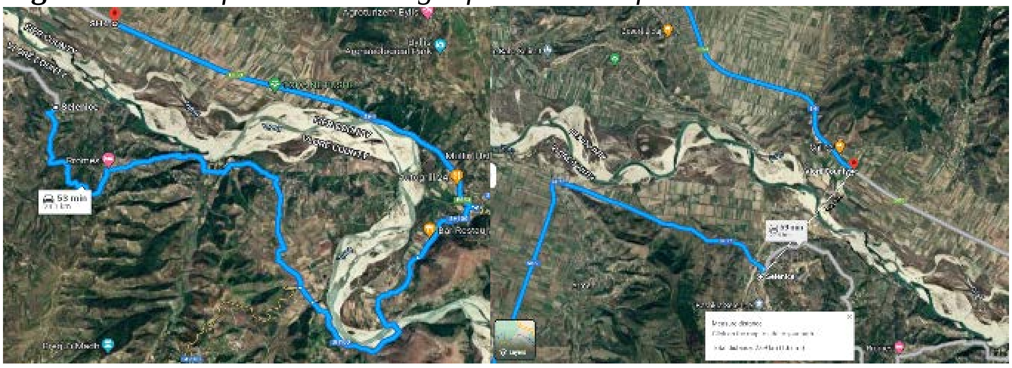
As was previously said, generative design enables a workflow that is more tightly connected between the designer/engineer and computer (Figure 2). In actuality, they both contribute to the final design.

Case Study

Bridges are intricate geometric constructions, and the many structural options typically show substantial geometric differences in their designs. In addition, with the new instruments developed, the construction industry frequently requires reducing the computational cost, shorter model runtime development and analysis, and little to no material waste in light of the environmental emergency. A low level reusing models in projects of a similar size is implied by the modeling complexity.

In order to achieve the aforementioned goals, the present paper suggests a generative technique to improve the bridge design process. This approach increases efficiency by lowering computational costs and modeling efforts. The methodology that follows uses a workflow to develop adaptable geometric models while introducing numerical and parameter correlations between each design parameter. As a result, by changing the parameter settings within the same model, new the elements of a bridge's geometric solutions can be produced via a generative development. Finally, the goal of the current work is to specify a modeling and analytic technique for a bridge project based on structural analysis, parametric development, and optimization. The outcomes can be used to better integrate the structure modeling in order to explore and develop high-probability designs complicated geometries and discover affordable solutions in the future.

Despite the fact that generative design can be applied in a variety of ways, the following is a study case of a connection bridge between the Fieri and Vlora County. Those two counties are being separated by Selenica river where there is no connection between the two existed highways. Taking into account that these areas are less than 3 km far from each-other, in this study there is conducted a generative analysis through the scripts of the grasshopper software and several plugins to give a solution to that. In order to get a wider view of the study case, there have been conducted through google map the measurement of the distance between two areas (Figure 3).

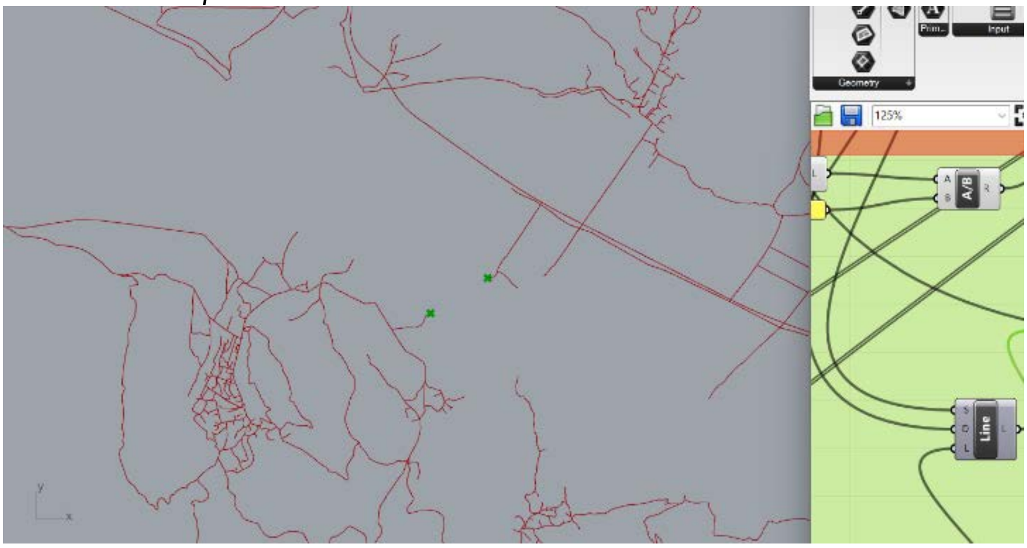
Figure 5. *Photos from the Existing Infrastructure of the Studied Area*

Source: Google Map, 2023.

Using a first script via Urbano plug-in of Grasshopper, it is conceived the converting infrastructure process into vectors. Practically, and programmatically, there is no difference between points and vectors. Both are lists-of-coordinates with a bunch of associated operations. In fact, plenty of programming platforms/ languages do not distinguish between points and vectors, and sometimes not even between vectors and colours.

However theoretically, and mathematically, they are very different entities and it is just easier to think with them if you keep them separate. As mentioned in several authors already, points are locations in space specified using a set of coordinates, whereas vectors are directions magnitudes in space, specified using a set of coordinate differences. Vectors are not geometry, and whenever we draw a vector in some specific place, we can only do so because we know, from the context, where that vector makes sense.

The process is following by finding the closest points between the parts of the infrastructure, which are being separated by Selenica River. In this way, it is very close to mind to make the solution by joining the two close points with a straight line by obtaining in regard a solution: a 600 meters long bridge that connects both sides. The visualized script is given in Figure 4.

Figure 6. *The Converting Infrastructure Process into Vectors and Finding the Closest Points of the Ground Terrain*

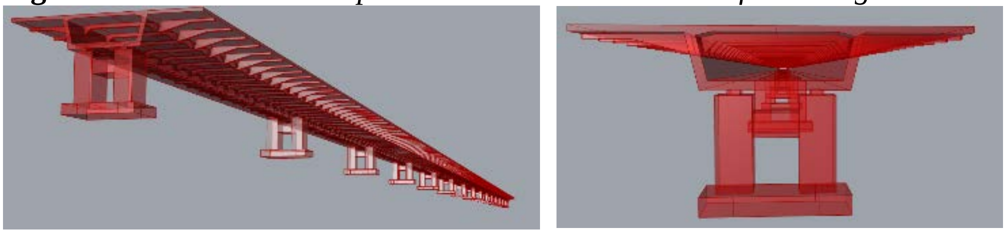


Source: Authors script, 2023.

Simpler Structural Modifications

Sometimes it is necessary to alter the construction of the road, possibly to integrate it with a nearby feature. The model can be adjusted with the robust tools, referred to as "Limits," to accommodate any auxiliary component. A linear element can be readily utilized to align the road model in both the horizontal and vertical directions once it has been captured as an axis. The study case is referred to a 600 metres long bridge with the technical parameters and the modelling bridge proposed for the study case is given in Figure 5.

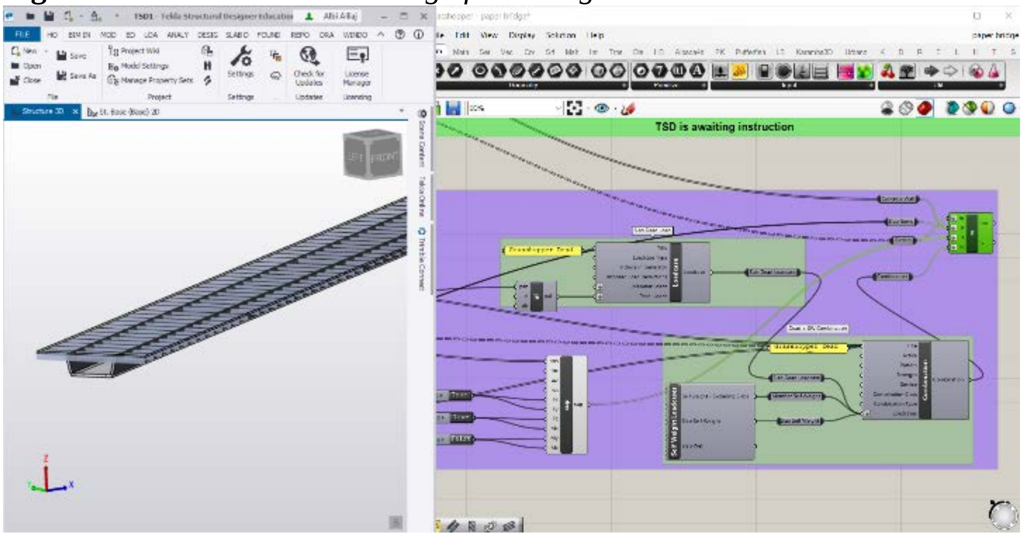
Figure 7. *The Technical Properties and the Cross-section of the Bridge*



Source: Authors design, 2023.

Parametric structural analysis of the deck is done utilizing Tekla Structural Designer Link plug-in inside Grasshopper3d. The first effort, where the girders are modelled with an identical thickness of 20 cm (Figure 6).

Figure 8. *The Structural Modelling of the Bridge*



Source: Authors script, 2023.

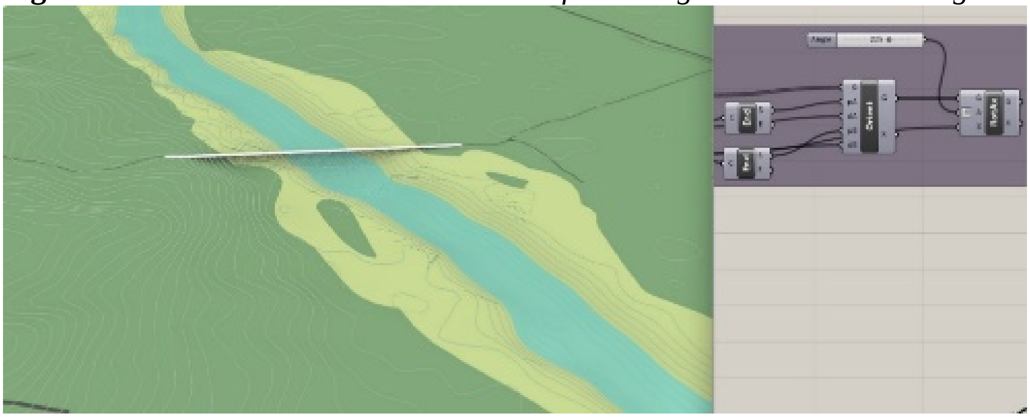
After that, were checked the girders. It resulted that most of the girders failed due to the loads considered (Figure 7).

Figure 9. *The First Structural Check of the Bridge*

Source: Authors script.

So, to prevent significant deformations and improve durability, it was simple to switch the model from uniform girders to variable cross-section girders and change the material of concrete class from C30/45 to C40/50. As a final step, it was conducted the verification of the deck structure due to major shear stresses, bending moments and deformations of the deck model of the bridge. As it can be seen from Figure 8, it has been obtained Schematic visualisation of the bridge model placed on the terrain generating by a full script (on the left).

Figure 10. *The Schematic 3D Visualization of the Bridge and the Surrounding*



Source: Authors, 2023.

Although a strong design application has been described in the study, it still needs some components before it can be applied to more bridge situations and be used by practical engineers. It is possible to further develop it, some of which are described below.

- More bridge design options should be included. Evaluating many design options is crucial early in the design process.
- For improved optimization and more accurate optimization outcomes, include surface topology in the definition.
- The members are presently picked from center to center. It follows by a procedure of cutting parts and alter the issues to permit more thorough modeling in, say, Tekla.
- Include a foundations study that takes into account the concrete, reinforcing, and geotechnical piles.
- Using machine learning to choose the best cross-sections for each set of elements.
- Expanding on CO_{2e} as an objective from the perspective of an LCA.
- Test out several optimization algorithms, including the Firefly algorithm, to see if it can get even better results.

Conclusions

In order to increase the effectiveness and efficiency of urban bridge infrastructure, the paper investigates the application of generative design and optimization approaches. It covers the problems with conventional bridge design and optimization techniques and how generative design can offer a better answer. A case study is presented that shows how generative design may be used to optimize a link bridge between two locations in an urban setting, producing a design that is both structurally effective and aesthetically pleasing.

In the framework of sustainable urban design, the use of generative design to optimize bridge infrastructure is also covered in this study. Additionally, an historical path of employing generative design in the context of infrastructure design and urban planning are discussed, along with how generative design might contribute to the development of more effective and sustainable bridge designs.

By adding more design factors and offering a fresh method for bridge optimization, the study was successful in further developing the initial parametric model. Due to the speed with which the design can be changed and the instantaneous display of the analysis, using parametric design to study alternative options for bridges could be highly helpful to designers.

Utilizing generative design is rapidly changing how manufacturers create the newest items. The need for project designs to perform better is one aspect of this. Another is for fresh, cutting-edge designs that provide their consumers exclusivity and customizability while utilizing the most recent manufacturing techniques. And because it is now more reasonably priced than ever before, designers can quickly and efficiently develop thousands of designs in a fraction of the time it would take to do so using a more traditional method.

Even while the definition of a parametric model is pretty solid as it is, it can still be improved. On the other hand, it offers the opportunity to be utilized in a real project in the early stages of design to really appreciate the benefits of the script. The first is whether the script for the following design stage may be utilized as a template (transferring it to Tekla), and the second is the degree of cost estimation accuracy in relation to the final design. Generatively developed products in the future will be networked, measuring information and utilizing it to train the algorithms and increase the technology's effectiveness. Future will be more intriguing than ever as this new generative design and all instruments used take center stage in contemporary design.

Data Availability Statement

The authors confirm that the data supporting the findings of this study are available within the article and its Supplementary material. Raw data that support the findings of this study are available from the corresponding author, upon reasonable request.

References

Autodesk, Inc. *What is Generative Design.* 2018. Available at: https://www.autodesk.com/solutions/generative-design.

Bader J, Zitzler E (2011) Hype: An algorithm for fast hypervolume-based many-objective optimization. *Evolutionary Computation* 19(1): 45–76.

Balachandran M, Gero J (1987) A knowledge-based approach to mathematical design modeling and optimization. *Engineering Optimization+ A35* 12(2): 91–115.

Bleiberg T, Shaviv E (2007) Optimisation for enhancing collaborative design. In *Proceedings of the Building Simulation*.

Broughton T, Tan A, Coates P (1997) The use of genetic programming in exploring 3d design worlds. In *CAAD Futures.* Springer.

Cao K, Batty M, Huang B, Liu Y, Yu L, Chen J (2011) Spatial multi-objective land use optimization: extensions to the non-dominated sorting genetic algorithm-ii. *International Journal of Geographical Information Science* 25(12): 1949–1969.

Coates P, Appels T, Simon C, Derix C (2001) Current work at ceca. In *Proceedings of the 4th Generative Art Conference* (GA2001).

Colton S, Cook M, Hepworth R, Pease A (2014) On acid drops and teardrops: Observer issues in computational creativity. In *Proceedings of the 7th AISB Symposium on Computing and Philosophy*.

Cui J, Tang M (2012) Presenting 3D Shape Grammars in a Generative Product Design System. In j Gero (ed.), *Proceedings of the 5th International Conference on Design Computing and Cognition*, 7-9 June 2012. USA. Texas A&M University.

Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation* 6(2): 182–197.

Derix C (2009) In-between architecture computation. *International Journal of Architectural Computing* 7(4): 565–585.

Gero JS (1975) Architectural optimization-a review. *Engineering Optimization* 1(3): 189–199.

Gero J, Radford A (1984) *The place of multi-criteria optimization in design. Design Theory and Practice*. London: The Design Council.

Hillier B (2007) *Space is the machine: a configurational theory of architecture*. Space Syntax.

Koenig R, Miao Y, Knecht K, Aichinger A, Konieva K (2020) *Integrating urban analysis, generative design, and evolutionary optimization for solving urban design problems*. Environment and Planning B.

Krish S (2011) A practical generative design method. *Computer Aided Design* 43: 88–100.

Maver TW (1995) Caad's seven deadly sins. In *Sixth International Conference on Computer-Aided Architectural Design Futures*, 21–22. 1995.

McCormack J, Dorin A, Innocent T (2004) Generative Design: A paradigm for design research. In J Redmond, et al. (eds.), *Proceedings of Futureground, Design Research Society*. Melbourne.

Miao Y, Koenig R, Knecht K (2020) The Development of Optimization Methods in Generative Urban Design: A Review. In *SimAUD: Symposium on Simulation for Architecture & Urban Design*. Vienna, Austria.

Monizza GP, Bendetti C, Matt DT (2018) Parametric and generative design techniques in mass-production environments as effective enablers of industry 4.0 approaches in the building industry. *Automation in Construction* 92: 270–285.

Pohl J, Myers L, Chapman A (1990) *Icads; an intelligent computer-aided design environment*. ASHRAE Transactions (American Society of Heating, Refrigerating and Air-Conditioning Engineers); (United States) 96, CONF-9006117.

Radford AD, Gero JS (1987) *Design by optimization in architecture, building, and construction*. John Wiley & Sons, Inc.

Ramanauskas R (2020) *Thoughts on the future of Generative Design in AEC from an engineering perspective*. BLOG on AEC Innovation. Available at: https://www.invokeshift.com/thoughts-on-the-future-of-generative-design-in-aec-from-an-engineering-perspective/.

Schmitt G (2004) The impact of computer aided architectural design on physical reality. *International Journal of Architectural Computing* 2(1): 31–41.

Schmitt G, Oechslin W (1992) *Computer aided architectural design futures*. In *CAAD Futures*, volume 91.

Simon HA (1975) Style in design. *Spatial synthesis incomputer-aided building design* 9: 287–309.

Simon HA (2019) *The sciences of the artificial*. MIT Press.
Sindhya K, Miettinen K, Deb K (2012) A hybrid framework for evolutionary multi-objective optimization. *IEEE Transactions on Evolutionary Computation* 17(4): 495–511.
Singh V, Gu N (2012) Towards an integrated generative design framework. *Design Studies* 33(2): 185–207.
Stouffs R, Rafiq Y (2015) Generative and evolutionary design exploration. *AI EDAM* 29(4): 329–331.
Turrin M, Von Buelow P, Stouffs R (2011) Design explorations of performance driven geometry in architectural design using parametric modeling and genetic algorithms. *Advanced Engineering Informatics* 25(4): 656–675.
Wetter M, Wright J (2004) A comparison of deterministic and probabilistic optimization algorithms for nonsmooth simulation-based optimization. *Building and Environment* 39(8): 989–999.
Widdowson W (1971) *Architecture as art: aphenomenological theory of architectural esthetics*. University of Pennsylvania.
Zitzler E, Laumanns M, Thiele L (2001) Spea2: Improving the strength pareto evolutionary algorithm. In *TIK-report 103*.

Athens Journal of Sciences- Volume 11, Issue 2, June 2024 – Pages 125-150

Infrastructures of Large-Scale Geothermal Energy Projects in Kenya: Materialization, Generativity, and Socio-Economic Development Linkages

By Chigozie Nweke-Eze[*]

The linkages between infrastructures and socio-economic development have become increasingly complex and varied in transdisciplinary human science scholarship. In the Global South context in particular, these linkages entail unusual geographies of diffusion that defies many easy narratives. Using the case of geothermal energy projects in Kenya, this article explores the materialization and generativity of infrastructures in large-scale projects and their complex linkages to socio-economic development. In so doing, the paper shows how the delivery of 'core' infrastructure projects enable the provision of 'other' infrastructures – 'required' and 'generated' infrastructures, all of which entail different socio-economic development linkages for different interest groups at national and local community levels. In exploring these processes, the paper engaged with multi-disciplinary scholarship on the materialization and generativity of infrastructures and their variegated and multifaceted linkages to socio-economic development. A methodological combination of expert and informal interviews, document analysis, and project-sites observations form the basis of our analysis.

Keywords: *infrastructures, large-scale geothermal energy projects, materialization, generativity, socio-economic linkages, Global South, Kenya*

Introduction

Infrastructures are apparatuses such as dams, highways, geothermal plants, canals, airports, and harbors, in energy, transport communication and water sectors of an economy or society, which enable other things to happen (Star 1999). In his influential review essay, Larkin (2013) further describes infrastructures as "built networks that facilitate the flow of goods, people, or ideas and allow for their exchange over space" (p. 328). For him, the "peculiar ontology" of infrastructure "lies in the facts that they are things and the relation between things" (Larkin 2013 p. 329). Following these lines of thinking leaves us with the understanding of infrastructures as a critical and necessary element for rapid socio-economic transformation. Yet, the existence of infrastructures or attempts to create them have generated critical debates on their potential causes of undesirable processes and outcomes such as human dispossessions, displacements, environmental degradation, and global warming (Beevers et al. 2012, Campbell et al. 2017, Divine et al. 2017). This paradox has increasingly become a subject of inquiry in interdisciplinary human science scholarship as many developing countries in the global south

[*]Researcher, Institute of Geography, University of Bonn, Germany.

increasingly "turn to infrastructure" with increasing mix of actors (Glass et al. 2019, Addie et al. 2020).

Infrastructures materialize through several complex processes of configurations, involving multifaceted actors with different interests and intents. The paper calls this process 'materialization of infrastructure'. The materialization of infrastructures also sets off other regeneration processes which lead to the materialization of other infrastructures for both intended and unintended purposes and consequences, which can go beyond the agency of the original developers (also see Silver 2014, Maringanti and Jonnalagadda 2015). The paper calls this the 'generativity of infrastructure'. These processes of infrastructure materialization and generation typically hold implications for broader socio-economic development in the spaces of their existence, through several linkages.

Using the case of the development of large-scale geothermal energy projects in Kenya, this paper explores the materialization and generativity of infrastructures of large-scale renewable projects and their complex socio-economic linkages in the developing countries context. The paper argues that a more complete appraisal of the socio-economic impacts of large-scale renewable projects should prelude a process-tracing analysis of their materialization and generativity potentials. It demonstrates this argument by showing how the materialization of large-scale geothermal energy infrastructures ('core' infrastructures) generates other infrastructures ('required' and 'generated' infrastructures), all of which, when considered as a whole, have multifaceted socio-economic implications, and impacts for interest groups at national and local levels. By so doing, this study responds to the growing calls to situate and understand infrastructure provisions in the realities faced by many countries in the Global South (Jaglin 2015, Coutard and Rutherford 2015). Empirical analysis of large-scale infrastructure projects in the Global South begs for wider thinking of the complexity and dynamism of infrastructure configurations, which challenges the predominant binary notion of their materialization and impacts (Lawhon et al. 2014, Silver 2014, Greiner 2016, Lawhon et al. 2018, Chambers 2019, Barry 2020).

The paper continues in the next section by discussing the infrastructures, their materialization, generativity, intents, interests, and their socio-economic development nexus. Afterwards, it presents the methodology and the cases of the study projects. Based on these cases, it goes on in subsequent section to analyze and discuss the materialization and generativity of geothermal infrastructure projects and their complex and differentiated socio-economic development interests and linkages at national and local community levels. The paper concludes by summarizing its findings and presenting its implications for socio-economic impacts analysis and appraisal of large-scale infrastructure projects.

Conceptual Framework

Materialization and Generativity Potentials of Infrastructures

The materialization of large-scale projects consists of a combination of infrastructure artefacts with generativity potentials to necessitate or enable the creation of other infrastructures in a networked configuration (Barry 2020). Heterogenous Infrastructure Configurations (HIC) formulated by Lawhon et al. (2018) provides an analytical lens which serves as a starting point in understanding these networked configurations. The HIC analyses infrastructure artefacts "not as individual objects but as parts of geographically spread socio-material configurations: configurations which might involve many different kinds of technologies, relations, capacities and operations, entailing different risks and power relationships" (Lawhon et al. 2018 p. 722). In doing so, Lawhon et al. (2018) push thinking around infrastructures to better consider and incorporate the numerous other complexities embedded within infrastructure construction, including stakeholder interests, thereby allowing for a distinguishing or separating infrastructural artifacts from one another, based on their interests and rationale for materialization and generativity. In this sense, infrastructures of large-scale projects are therefore not independent apparatuses but are often geographically embedded and networked in wider socio-material configurations of relations and operations, possibly in network with other technical and social infrastructures, with socio-economic and political implications (Silver 2014, Chambers 2019, Thekdi and Chatterjee 2019).

Before Lawhon et al. (2018), existing accounts attempted to frame the complex materialization and generativity potentials of infrastructures as hybrid and mixtures (Furlong 2020, Larkin 2008), continuous and incremental (Silver 2014, Maringanti and Jonnalagadda 2015), post-networked (Coutard and Rutherford 2011, Monstadt and Schramm 2017), as well as people-centered and lived (Graham and McFarlane 2014, Simone 2004, Scott 1998). The hybridity of infrastructures materialization and generativity reflects in the diverse and different ways in which infrastructure artefacts connects and embeds into existing infrastructure geographies, sometimes causing the creation of other new infrastructures (De Boeck and Baloji 2016, Kimari and Ernstson 2020). Although similar literature focuses on the spread of networked infrastructure, Meehan (2014) suggests the consideration of 'informal' infrastructures which can emerge in large-scale projects, and which often serve as conduits outside of state control. These networked infrastructures often inspire new possibilities for social collective organizing, ownership and power relations as well as generating new platforms for engagements outside of the state, which may or may not be initially intended (Schouten and Mathenge 2010, Ernstson et al. 2014, Silver 2014).

Infrastructures are also continuous and incremental in the sense that it involves constant socio-material production, maintenance, expansion and reconstruction (Silver 2014, Coutard and Rutherford 2015, Maringanti and Jonnalagadda 2015), with diverse involvement of people as actors in shaping its constitution and determining its generativity in mutual constitution – leading some authors to argue for the wider notion of infrastructure that includes 'people as infrastructure' (Simone 2004, Anand 2011, Larkin 2013, McFarlane and Silver 2017). These processes involve a wide range of actors at public and private, local and international, formal and informal levels, consisting of project developers, investors, of local entrepreneurs, grassroots social movements, international NGOs, and individual community members, each with different interests, motives, incentives, and perceptions (Lindell 2008, Pieterse 2019, Cirolia 2020). These increasing and diversified involvement of actors in infrastructure provision and the resulting generativity which they are increasingly creating, have inspired works that seek to show how infrastructures have become layered by additional and partial infrastructures, with different other uses, coverages, logics, and ownerships (Anand 2011, Chattopadhyay 2012, Graham and McFarlane 2014, Silver and Marvin 2017).

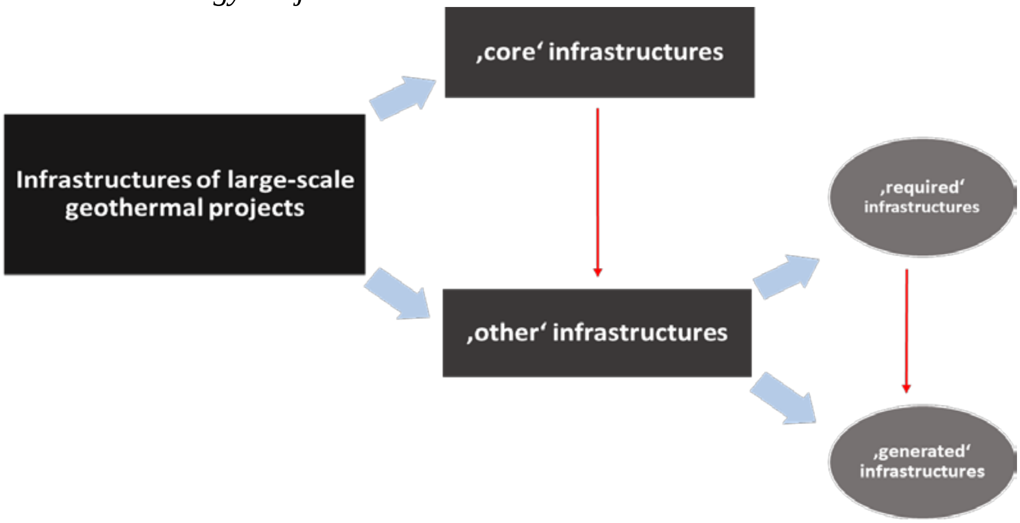
Materialization and Generativity of Large-scale Infrastructure Projects

Intended large-scale infrastructure projects become generative in the process of their materialization, allowing for the construction of other technical and social infrastructures in project-host communities. Infrastructures of large-scale projects primarily materialize in two forms ‘core’ and ‘other’ infrastructures. It starts with core infrastructures, which are the actual intended infrastructures, made up of the geothermal plants and machineries. It then goes on to show how the construction of these core infrastructures both enables and necessitates the construction of other new or additional infrastructures – ‘required’ and ‘generated’ infrastructures. ‘Required’ infrastructures are additional technical infrastructures, which are provided to enable the construction of the core infrastructure projects. Such infrastructures usually consist of access roads and large water supply and storage systems. As the project proceeds, these ‘required’ infrastructures then further enable the provision of other ‘generated’ technical and social infrastructures – network roads, water abstraction points, schools, hospitals, housing, which are often provided in form of corporate social responsibility (CSR) and as resettlement plans for Development Induced Displaced Persons (DIDPs). These ‘generated’ infrastructures would not have been provided[1] if the ‘required’ infrastructures were not initially provided. In general, the required and the generated infrastructures do not only enable the construction of core infrastructures, but they also exist to ensure their continuous functionality.

These materialization and generativity processes (Figure 1) reveal how infrastructures assume lives of their own and catalyze the materialization of further infrastructures, which have a multiplier effects on social-economic development patterns in project host-communities, especially in the peripheral and marginalized geographies where infrastructures are already scarce.

[1]Or would have at least been very difficult to provide or take a long time to be provided.

Figure 1. *Materialization and Generativity of Infrastructures in Large-scale Geothermal Energy Projects*



Source: Author.

Intents and Interests in Infrastructure Materialization and Generativity

The materialization and generativity of infrastructures in large-scale projects reflect and are conditioned by a combination of intents of diverse actors at international, national and community levels (Cirolia 2020, Nweke-Eze and Kioko 2021). Infrastructure rush in the Global South is fueled by the infrastructure-development nexus understanding, which emphasizes the importance of industrialization particularly through infrastructure as the key to economic growth and development (Cooper 1996, Luiz 2010). Investments for development infrastructures are, however, currently limited in the Global South, leading to greater push to attract more infrastructure investments from new classes global funders (Terrefe 2020, Van Noorloos and Kloosterboer 2018). These realities have contributed to the widening of the scope and scale of interests and intents to include the geo-political and economic interests of fund providers and financiers (Goodfellow 2020, Klagge and Nweke-Eze 2020).

At the same time, large-scale projects are associated with intents and interests at national and community levels. National governments interests in infrastructure provision are encapsulated in their intent to centralize power or to display the power of the state to foster national development and to delivery on national promises (Ballard and Rubin 2017, Cirolia and Smit 2017). Regardless of intent and interest, some of the infrastructure investments in the Global South have proven to be poorly coordinated leading to debt traps which result in dangerous continuities of macro-economic quagmires (Banerjee et al. 2008, Foster and Briceño-Garmendia 2010, Furlong 2020). At the community level where infrastructure projects are constructed, interests and intents are mainly directed towards meeting socio-economic requirements, while conserving the environment (Nweke-Eze and Kioko 2020). In some cases, community leaders have been shown to have vested interests in large-scale infrastructure projects, with the power to oppose and obstruct state provision of infrastructure and initiatives (Arrobbio et al. 2014, Klagge et al. 2020, Greiner et al. 2021).

Infrastructures and their Complex Linkages to Socio-economic Development

Studies have shown that the degree of development linkages of infrastructures depend on specific geographies, timing, and politics (Edwards 2002, Straub 2011, Howe et al. 2015, Anand et al. 2018, Furlong 2020). The benefits from infrastructures can be significant and vary depending on specific local contexts (Turner 2018, Weinhold and Reis 2008). Constructing new infrastructures or improving existing ones can increase access to new markets by of helping rural farmer access urban markets, increase prices of their products and make more profits; as well as increase access to social and institutional infrastructures such as schools, hospitals (Jacoby 2000, Mu and van de Walle 2011, Aggarwal 2018). However, the positive impacts of new roads can be heavily outweighed by other socio-economic livelihood losses, bio-diversity disruptions, and environmental damages (Foley et al. 2007, Mandle et al. 2015, Beevers et al. 2012). For instance, in certain local contexts, new infrastructures can adversely affect access to water for domestic purposes or fishermen who depend on the water bodies for their socio-economic livelihoods (Appiah et al. 2017).

The extent of positive impacts of infrastructure on development in a particular country also depends on what Calderon et al. (2011), Estache and Garsous (2012), Garsous (2012) and Estache and Wren-Lewis (2011) refer to as the "the development stage" of a country. The more developed a country is, the higher its infrastructure stock and hence the lower the payoff from additional investment, unless it aims at addressing a major bottleneck or introducing a major technological improvement (Estache and Garsous 2012, Garsous 2012). On the other hand, the less developed a country is, the more significance is the impact of an additional infrastructure (Estache and Garsous 2012, Garsous 2012). These literatures, however, also note that some infrastructure projects, such as energy and transport infrastructures, do have positive impacts regardless the development stage of the country (Estache and Wren-Lewis 2009, Estache and Garsous 2012).

Studies have shown that the time-period over which the impact is assessed also matters. The significance of the positive impact of infrastructures from the 1950s to the 1980s were more prominent than after the 1980s (Estache and Fay 2010). Studies that observe infrastructure impacts over longer time-periods were more likely to observe more significant positive impacts (Albala-Bertrand and Mamatzakis 2004, Estache and Fay 2010) – this has been attributed to the long payback period of most infrastructures (Estache and Garsous 2012). The degree of impact an infrastructure may have on socio-economic development also depends on the type of infrastructure (Dethier et al. 2008, Estache and Garsous 2012). Most findings show that direct-impact infrastructures, such as energy and information and communication technology (ICT) infrastructures tend to have higher positive significance on development indices than other more indirect-impact infrastructures such as water and sanitation infrastructure, which often depend on other infrastructures (example, energy infrastructures) to function (Garsous 2012).

Large-scale infrastructure projects often have far-reaching socio-economic impacts, often extending beyond the immediate spatiality of the project site, into nearby and further spaces, with varying temporal (short, medium, and long-term or even permanent) effects (Batey et al. 1993, Korytárováa and Hromádkaa 2014). Studies such as Enns and Bersaglio (2020) and Bryceson et al. (2008), contend that infrastructures connect to socio-economic development in a selective and uneven manner – stating with empirical evidence that certain infrastructures have increased socio-economic development for some, while at the same time worsening socio-economic development and welfare for others.

Infrastructures are only useful to the degree they help to facilitate activities. Such facilitating activities of their provision, accessibility, reliability, scale, durability, and maintenance, allows us to differentiate the degree and extent of impacts of infrastructures in different geographical contexts (Amin 2006, Hall et al. 2013, Talen 2019). As Bryceson et al. (2008) argue, infrastructures in themselves are blunt instruments which must co-exist with certain other enabling conditions and means to effectively translate or contribute to socio-economic development. The variegated impacts created by the differentiated quality of infrastructure facilitating activities has led to non-uniform outcomes of infrastructure provision (Lawhon et al. 2018).

Infrastructures and their Multifaceted Socio-economic Development Impacts

Large-scale infrastructures such as energy projects (electricity generation, transmission, and distribution systems), water projects (pumping, boreholes and sanitary systems), transportation projects (roads, railways, ports, pipelines), information and communication technology projects (broadband masts, telecommunication systems) have long been part and parcel of human socio-economic life. The development of these infrastructures is often connected to and/or justified in the mainstream development circles by grand narratives of development/underdevelopment, as conditions in which prosperity of nations are bound (Kanai and Schindler 2019). This infrastructure-development nexus has come to dominate national and international development policy agenda, subsequently leading to a surge of interest in infrastructural development, investments, and financing spear-headed by state and regional governments and supported by several old and new, international, and regional development institutions, multi-donor, and climate agencies (Boyer 2019, Howe 2019, Klagge and Nweke-Eze 2020).

The ideology and perspective on infrastructure-development nexus have been subject to discourse, starting from Arrow and Kurz (1970) and Aschauer (1989). Since then, many other (inter- and multi-) disciplinary studies have begun to analyze, discuss and debate the subject matter. Generally, the findings of these studies are bifurcated. Many studies from national economic growth and development perspectives predominantly highlight the positive impacts of infrastructures based on macro-economic indices. These studies generally report that increase or improvement of infrastructures brings about positive impacts on several socio-economic and development indicators, including long-run economic growth, international trade enhancement, productivity and efficiency, economic development; poverty alleviation and the achievement of the Millennium Development Goals (MDGs) (Asher and Novosad 2020). Exemplary for this literature are studies on impacts of infrastructure are studies by Easterly and Rebelo (1993) and by the World Bank (1994) who conducted global, multi-country research in both the Global North and South; studies by Seethepalli et al. (2008), Straub (2008) and Straub and Terada-Hagiwara (2011) who focus on East Asia; and Calderon and Serven (2008) Calderon and Chong (2009) who conducted research in Sub-Saharan-Africa.

In contrast, however, studies researching from mainly local community development and bio-diversity perspectives report mainly negative impacts of infrastructure projects on biodiversity and environment (Trombulak and Frissell 2001, Laurance et al. 2006, Coffin 2007, Campbel et al. 2017), and their disrupting effects on indigenous people's livelihoods (Kenley et al. 2014, Collier et al. 2015, Barker et al. 2021). They report incidences of human-vehicle collision and accidents, animal-vehicle collision and accidents, noise pollution during construction of project or usage of infrastructure projects such as roads, restriction of movements, reproduction patterns and other disruptions of wildlife, increased spreading of invasive plants, landscape disasters such as landslides and erosions, and increased hunting, poaching, deforestation and other human-wildlife interferences.

Methodology and Case Studies

Methodology

The analyses and discussion of the study is based on expert interviews (2018-2020) with interview partners who work at different levels of government[2] (MoE[3], the National Treasury, County and national commissioners), and in energy-related and other state agencies (ERC, GDC, KenGen, KETRACO, KWS, NLC)[4]. It also features interview partners in development finance institutions (AFD, AfDB, EIB, KfW, TDB, USAID)[5], in private consulting firms (Tetra-Tech, GeoHydro Energy Consultants Limited) and in an energy research institute (GETRI)[6]. In addition, the study analyses are also based on analysis of project reports, several project sites visits and observations as well as informal interviews with project staff and local community members and in the projects host communities (2018-2020).

[2]National, local and, since 2013, county levels – following the devolution of government functions in Kenya.

[3]Ministry of Energy.

[4]Energy Regulatory Commission, Geothermal Development Company, Kenya Electricity Generating Company, Kenya Wildlife Service, National Land Commission.

[5]Agence Française de Développement, African Development Bank, European Investment Bank, Kreditanstalt für Wiederaufbau, Trade and Development Bank, United States Agency for International Development

[6]Geothermal Energy Research and Training Institute.

Case Studies: Large-scale Geothermal Energy Projects in Kenya

In this section, we discuss the three geothermal projects in Kenya that constitute our case study, namely: Olkaria, Menengai and Baringo-Silali (see Figure 1). The Olkaria project is the oldest and most advanced of the projects. It already generates about 623MW of electricity (KenGen interview 2019-2021, Figure 2). This is followed by the Menengai project, which was as at the period of fieldwork in 2019-2021, power plant construction by independent power producers (IPPs) for the first generation of 105MW of electricity is being planned (Figure 2). At the same period, the Baringo-Silali project was still in project exploration and test drilling stages (Figure 2).

Olkaria geothermal project is located in a semi-peripheral area of Naivasha town, Nakuru county, partly in Hell's Gate National Park (a touristic Wildlife Reserve) and in partly on the homeland of Maasai people[7]. Menengai geothermal project is located in semi-peripheral area of Nakuru town also in Nakuru county, with most parts within the Menengai Crater in Bahati sub-county[8] and a smaller part encroaching in previously privately-owned land[9] (NLC interviews 2019). Nakuru county spans an area of 2,325.8 sq km with a population of 1,503,325 according to the 2009 census. Communities in both Naivasha and Nakuru town mainly engage in trading and farming. In contrast, the Baringo-Silali project (consisting of Paka, Korosi and Silali) is located in the peripheral, semi-arid Baringo county in Kenya, on communal land (NLC interview 2019). Baringo covers an area of 11,015.32 sq km with a population of 555,561 according to the Kenya Census data 2009.The dominant ethnic groups are the Pokots, Tugens, Endorois and Ilchamus. These communities mainly keep livestock, although the people living in the highlands practice farming.

The several components of Olkaria Geothermal project are majorly developed by KenGen[10] as well as by OrPower4 Inc[11] and Oserian Flowers Ltd[12]; while Menengai and Baringo geothermal energy projects are developed by GDC[13]. The three geothermal energy projects received technical and debt and grants financing support from development financial institutions such as AfDB, TDB, EIB, KfW, AFD, JICA[14], USAID and the World Bank, as well as by climate agencies such as SREP[15] and the GEF[16], at various stages of the projects' development (GDC, KenGen, National Treasury, DFI interviews 2019).

[7]KenGen (the project developer) resettled the Massai people previously living in these areas in order for the project to carry on (KWS, NLC & KenGen interviews 2019).

[8]Bahati sub-county is located close to an urban area where lands are predominantly privately owned (NLC interview 2019).

[9]The land was bought from their private landowners for the project to continue (NLC interviews 2019).

[10]A limited liability company with 70% Kenyan government shareholding.

[11]Private-owned company and an independent power producer licensed to develop the third component of the project (Olkaria III) (ERC interview 2019).

[12]A private company growing flowers for export, with its own geothermal power plant.

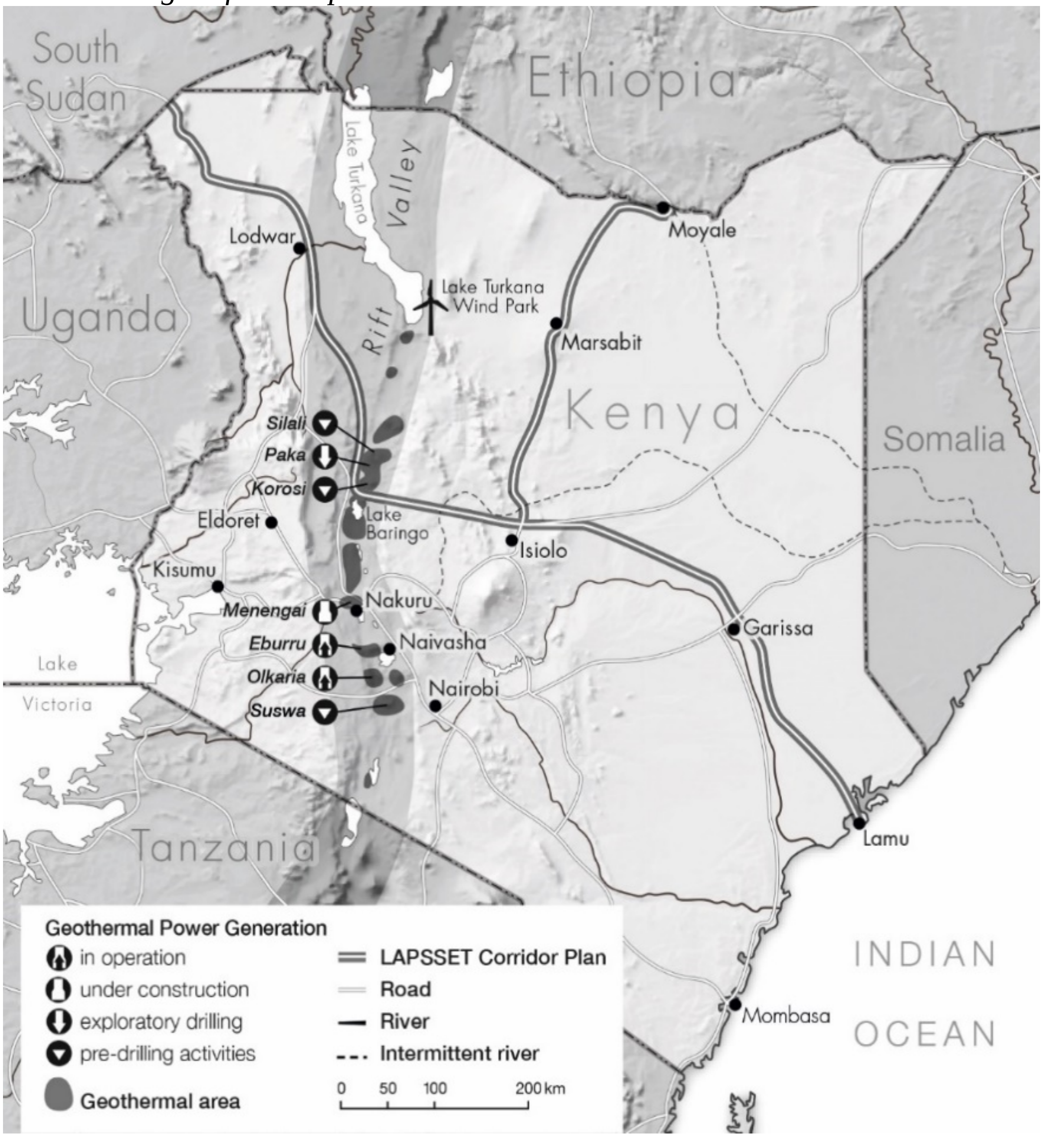
[13]A government owned special purposed vehicle established in 2008 with the mandate to explore and develop geothermal field in the country.

[14]Japan International Cooperation Agency.

[15]Scale-up Renewable Energy Program.

Although geothermal energy projects differ depending on their location, they generally go through similar stages and processes before their commissioning and operation. Preliminary surveys and exploration, test drilling and reservoir confirmation and feasibility studies, are first carried out to confirm the viability of the project development. This is then followed by actual site development, which then leads to start-up and commissioning of the project.

Figure 2. *Map Showing Geothermal Fields and Sites in Kenya, their Locations, and their Stages of Development*



Source: Klagge AND Nweke-Eze (2020).

[16]Global Environment Facility.

Results and Discussion

Infrastructures in Large-scale Geothermal Projects: Materialization and Generativity

Using the case of large-scale geothermal projects in Kenya, this section shows and discusses how intended large-scale infrastructure projects become generative in the process of their materialization, allowing for the construction of other technical and social infrastructures in project-host communities. It reveals how infrastructures of large-scale projects primarily materialize in two forms 'core' and 'other' infrastructures. The following three sub-sections further discuss the materialization of these infrastructures in categories of their generativity potentials.

<u>Geothermal Plants and Machineries as 'Core' Infrastructures</u>

Generally, the geothermal power plants use steam obtained from geothermal reservoirs to generate electricity. Prior to commencement of the work for the power station, production and injection wells are drilled at the appropriate locations to bring this geothermal energy up to the surface (GDC 2010). A mixture of steam and water is then collected from the production well, which are then separated using the steam separators. The steam is used to operate turbines which powers the generators, hence, generating electricity. The condensed steam and the water collected from the production well are injected back into the reservoir through the injection well (GDC 2010).

Other than the above-described power plants, other facilities in the geothermal power project sites are called Steam field Above Ground System (SAGS) (GETRI interview 2019). They consist of the steam pipelines, brine/condensate pipelines, separators, scrubbers, and the rock mufflers (GETRI interviews 2019). Geothermal steam & fluid from production wells is piped downhill from the separators as two-phase flow (GETRI interviews 2019). The pipelines are made of carbon of robust inches (GDC 2010, Fieldwork 2019). First, there are pipelines from each well pad to separator (GETRI interviews 2019). These are then followed by the steam pipelines from the separator to the power station, the brine pipeline from the separator to each injection well pad, and the condensate pipeline from the cooling water piping to the injection well pad (GETRI interviews 2019). Necessary pipe loops are provided on those pipelines to absorb thermal expansion (GETRI interviews 2019).

The cyclone-type separators are used to separate steam from two-phase liquid coming from production wells (GETRI interviews 2019). Steam goes to power station while the brine goes to injection wells (Project sites observations 2019). Scrubbers of corrugate type are provided just before the power station to eliminate further moisture (GETRI interviews 2019). Surplus steam is released to the atmosphere through vent valves (Project sites observations 2019). Rock mufflers are provided near the separator station to reduce the noise level of the released steam (GETRI interviews 2019).

<u>Access Roads and Water Systems as 'Required' Infrastructures</u>

The construction and operation of these 'core' infrastructures necessitate the delivery of 'required' infrastructures, namely: access roads and water pumping and storage systems. The access roads, as the name implies, provide access to the project site, and connect the core project sites to stand-by water system and the equipment-offloading storage sites (Project sites observations 2019, 2020; GDC interviews 2019). These access roads are necessary for transporting heavy well exploration and drilling equipment such as exploration and drilling gears and pumps, drilling rigs, hydraulic excavators with large diameters and thickness; the Steamfield Above Ground System (SAGS) as well as other materials such as diesel fuel, cement and concrete and (in some cases) water with bulk mass; into the project field or site (Project sites observations 2019, 2020; GDC interviews 2019).

The access roads are provided either by improving the capacity of already existing roads through expansion, or by constructing entirely new ones, usually in marginalized peripheral areas where there were no prior existing roads leading to the project sites (GDC 2010, 2013, 2019). Access roads for the projects are fortified with several layers of gravels before surfacing in other to withstand the frequent movement of heavy vehicles, equipment and materials, over-time (GDC 2010, 2013, 2019; Project sites visits 2019, 2020).

Geothermal project construction will typically not materialize on site without the delivery of water pumping and storage systems, which come in different scales depending on the size of the project. The pumped water is used for testing steam and for mixing materials during the construction phases of the project (Project sites observations, 2019). Other than for the development of the project, water also plays an integral role of steam generation in flash and binary geothermal power plants[17] (GDC and KenGen interviews 2019). During the operation of the geothermal power plant, water is used in both high- and low-pressured form to generate steam, which is used to drive the geothermal turbine for the generation of electricity. The pumped water is sourced from nearby water bodies, using diesel-fuel-power generators and through laid-pipes, into large water storage systems (Project sites observations, 2019; GDC 2010, 2013, 2019). Stored water from the storage tanks is then pumped or excavated through other pipes which connect the stored water systems to the project sites, when needed (Project sites observations, 2019; GDC 2010, 2013, 2019). Projects which are developed in areas that are far from water bodies, where construction of laid-pipe are non-feasible, often depend on large water-tank-vehicles which carry water over long distances using already existing or constructed access roads (Project sites observations, 2019; GDC interviews 2018). 'Required' infrastructures also include other infrastructures, such as temporary or sometimes permanent water and housing structures for project workers in host communities (Project sites observations, 2019; GDC interviews 2018).

[17]Most modern geothermal power plants are flash or binary. Binary geothermal power plants are said to be the power plants of the future (KenGen interviews 2019).

Other Technical and Social Infrastructures as 'Generated' Infrastructures

'Generated' infrastructures are the infrastructures that follow and because of the provision of the 'required' infrastructures, in the development of large-scale geothermal projects. These 'generated' infrastructures are provided in several forms: as extension of already existing required infrastructures, as Corporate Social Responsibility projects or activities, as part of resettlement schemes for project affected person (PAPs) (GDC 2010, 2013, 2019, Fieldwork 2019, 2020). These 'generated' infrastructures include technical and social infrastructures such as road networks, water abstraction points, and community housing structures like schools, hospitals, residential buildings, etc. (GDC 2010, 2013, 2019, Project sites observations, 2019, 2020).

Infrastructures of Large-scale Geothermal Projects: Socio-Economic Development Linkages and Interests

In this section, we use our case study of three different geothermal projects in Kenya to contextualize and illustrate the socio-economic development linkages of infrastructures in large-scale projects. As shown in section 3 (Figure 1), the projects are in different stages of their development, with Olkaria being the most advanced consisting of already existing plants (see Olkaria II in Figure 3), generating over 600MW of electricity. As at the time of the fieldwork, Menengai is at the final phases of development and was readying for fitting in steam capturing plants for electricity generation; while the Baringo-Silali block is in its preparatory stages of drilling and was recording its preliminary steam striking successes.

Infrastructures of Olkaria Geothermal Projects

The Olkaria geothermal plants, as 'core' infrastructure projects, were constructed primarily for the purpose of electricity provision at the national level, because of the centralized nature of the Kenyan national electrification plan (MoE interviews 2019, 2020). Therefore, although the geothermal electricity is generated at the local project-host community level, access to electricity in the community is determined in top-down decision framework originating from decisions and planning at the Kenyan Ministry of Energy (MoE interviews 2019, 2020).

Figure 3. *Aerial View of Olkaria II Geothermal Plants and SAGS*

Source: KenGen archives (2020).

The project development was preceded by the provision of access roads and water supply systems as initial required infrastructures. The 24 km Moi South Lake Road (MSLR)[18] had existed for a long time but had mostly remained in a bad condition. The planned development of geothermal projects and the existence of the Hell's Gate National Park and flower farms in the area, sparked the discussion for and eventually led to the tarmacking of the road (Kuiper 2019, KenGen and MoE interviews 2019). The tarred road is sporadically maintained and repaired by flower farms, hotels and several other NGOs operating in the area, with some contributions from KenGen (Kuiper 2019; KenGen interviews 2019). The access road was used for transportation of construction equipment and materials used for the construction and maintenance of the different components of the Olkaria geothermal project. The access road additionally provided right-of-way for the construction of transmission lines for the evacuation of generated electricity to the national grid (KenGen interviews 2019). The MSLR is the only paved class D road[19] in Naivasha district of Nakuru county so far, providing quicker access to the main Nairobi-Naivasha highway. This connectivity has enabled quicker transportation of farm produce from the project region, as well as increased access to social infrastructures in nearby towns (Ogola 2013, Fieldwork 2020). The road, however, also increased air and noise pollution from vehicles and increased the number of illegal and informal settlements in the area.

Other than access roads, water pumping, and storage systems were also constructed from water sourced from Lake Naivasha. The water systems were used as a drilling fluid during construction and for well-testing during construction stages of the project. The water systems are also maintained and utilized for pumping water for operating the Olkaria geothermal power systems (GIBB Africa 2009, Fieldwork 2019). The pumped water was then further purified by KenGen and piped for use by the surrounding Massai communities at several community water-points, as part of CSR (Ogola 2013, Fieldwork 2020). Four Massai villages[20] were resettled due to concerns for noise pollution and the emission of Hydrogen Sulphide gas (H_2S) at dangerous levels during the construction of the Olkaria IV project (Fieldwork 2019). The resettlement action plan (RAP) provided for the resettlement of the four villages as one entity with the provision of resettlement infrastructures including roads, pipe-borne water[21], electricity, houses, schools, health centres, lands and land title deeds[22], all of which cover a space of 1700 acres (KenGen interviews 2019, Schade 2017, pp. 13–14). There are however concerns over the efficiency and suitability of the resettlement scheme, as a result of massive records of dissatisfaction among many of the resettled community members (Schade 2017, Nweke-Eze and Adongo forthcoming).

---

[18]Code named D-323.

[19]Class D roads are secondary roads according to the classification of roads in Kenya

[20]The four villages were: Cultural Centre, OloNongot, OloSinyat and OloMayana Ndogo.

[21]5 water structures were constructed for the benefit of humans and livestock in the resettled communities as well as in Narasha, Maiella and Iseneto.

[22]The provision of land title deed was very significant in the resettlement process, as it was the first-time project affected persons (PAP) would become official landowners upon resettlement.

Infrastructures of Menengai Geothermal Projects

At the time of Menengai geothermal project development, the region surrounding the project site was already well serviced by a network of earth roads and all-weather roads, linking up the Nairobi-Kisumu Railway line and trans-Africa highway passing through the southern part of the area (GDC interviews 2019). The Menengai crater, which constitutes a major part of the project's site, had long been an attraction site for tourists and a site for excursion for school pupils and students. The already available access roads leading up to the project site were, however, widened to make it adequate for transporting heavy plant and equipment, personnel, and project supplies (GDC interviews 2019). New network roads connecting to these already existing access roads, were then constructed to further open access to the region for the host communities, for new business creation and expansion of existing ones.

In addition, the government-owned developer Geothermal Development Company (GDC) constructed a 20-million-liters water storage system for storing water sourced from Lake Naivasha. The stored water was used for cooling the power plants during drilling and for well-testing; and is further maintained for use to operating the power plants when they are constructed (GDC 2013, Project sites observations, 2019). As the Menengai geothermal project is located in the Menengai crater, there were no displacements of the communities in villages of Bahati sub-county (GDC 2013, Project sites observations 2019, GDC interviews 2019) (Figure 4). However, private farmers whose lands were acquired for road expansion and whose farmlands were affected by the passing of the power transmission lines were compensated in monetary terms (GDC & NLC interviews 2019, Fieldwork 2019).

Figure 4. *Menengai Geothermal Project Water Storage Systems*

Source: GDC archives (2020).

Infrastructures of Baringo-Silali Geothermal Projects

Unlike the Olkaria and Menengai geothermal projects located in Nakuru county – a semi-peripheral area with some existing infrastructures before the development of the projects, the Baringo-Silali geothermal project is in Baringo county – a peripheral and marginalized area of northern Kenya where infrastructures were scarce. For this reason, ample time was taken to build access roads, out of bare pathways, before the project developers were able to move plant machineries and equipment to the project site (GDC 2019, GDC interviews 2019). A 70km access roads were completed and more than 100km of existing roads were expanded and paved[23], creating a robust road network[24]. These roads are, however, not tarred (see Figure 5), leading to air pollutions (dusts) as heavy and light vehicles drive in high speed along the roads (Project site visits 2018, 2019). The construction of roads was followed by the construction of water pumping systems together with 4.5-million-litre water tanks for storing water sourced from Lake Baringo in Paka, Korosi and Silali (Project site visits 2018, 2019). The water pumping and storage systems were constructed for sourcing water for drilling and cooling activities during geothermal site development and will be maintained and utilized for operating the geothermal power plants at a later stage (GDC interviews, 2019, (Project site visits 2018, 2019).

Figure 5. *Aerial View of 'required' Infrastructures in Baringo-Silali Geothermal Project Site*

Source: GDC archives (2020).

Water from the storage tanks is purified and piped for domestic use in the community[25] through 20 newly commissioned watering points and water treatment

[23]The paved B4 road running upward-north through Marigat ending in Chemolingot.

[24]Paka – Silale; Kadingding – Korossi; Korossi – Lomuge; Naudo – Akwichatis; Chepungus – Kadokoi.

[25]For both humans and animals.

plants[26], as part of CSR (Project site visits 2018, 2019, GDC 2019). These watering points were however not initially planned; they were constructed upon the request of the host communities during negotiations (GDC interview 2019, Community members interviews 2019). Before the construction of the watering points, portable water was, for the meantime, periodically provided using large water-tank-vehicles, which carry water over long distances using already existing or constructed access roads (GDC interview 2019, Community members interviews 2019). GDC is also involved in further CSR activities in the project area (GDC 2019, Fieldwork 2019). It constructed an Early Childhood Development (ECD) classroom at Kibenos in the North Rift Valley and provided scholarships to needy students in the project area to attend universities, secondary and primary schools (GDC 2019, Fieldwork 2019). Since, there is no project displaced persons so far, there were no resettlement infrastructures in the development geothermal energy in the area (GDC 2019, Fieldwork 2019).

Figure 6. *Provided Community Water Point in Baringo, as Part of Corporate Social Responsibility*

Source: GDC archives (2020).

Table 1 summarizes the three categorizations of infrastructures in large-scale geothermal energy projects in Kenya; depicting their types, means of materialization and socio-economic development linkages.

[26]Kadingding, Mesori, Nakuórojang, Moinonin, Cherisan (Pump station I), Tuwo, Chepungus, Reong'o, Chemoril, Natan, Naudo, Angromit, Ponpon, Orus, Katungura, Kwokwototo, Nasorot, Korossi (tank site), Adomejong, Akwichatis. As the time of the fieldwork (2018-2019), some of the provided water points were still under construction at the commissioning, while some of the finished ones were not functioning at full capacity – lacking water at times.

Table 1. *Infrastructures of Large-Scale Geothermal Projects and their Socio-Economic Development Linkages*

Categorization of infrastructures in large-scale projects, based on their generativity	Infrastructure types	Materialization	Socio-economic development linkages and interests
‘core’ infrastructures	Power plants	Actual projects	Electricity provision, serving interests at the national level
	Steamfield Above Ground System (SAGS).		
‘required’ infrastructures	Access roads	Project development requirement	Access to project sites and to markets, serving interests at both national and community levels
	Water pumping and storage systems	Project development requirement	Water for construction and geothermal steam production, serving interests at project developers’ level. By generating electricity, the infrastructure ultimately serves interests at national levels
‘generated’ infrastructures	Network roads	Corporate Social Responsibilities (CSR) or community improvise	Market connections and mobility, serving interests at community levels
	Community water points		Water supply for domestic and agricultural use, serving interests at community levels
	School buildings	CSR, Resettlement schemes*	Education, serving interests at community levels
	Health centres		Health services, serving interests at community levels
	Housing	Resettlement schemes*	Modern shelter, serving interests at community levels (while and fulfilling resettlement criteria of the investors)

*The suitability and impact of these modern housing infrastructures, which were provided as part of the resettlement schemes, are however questioned and debated (Schade 2017, Nweke-Eze and Adongo in-print).

Differentiated Provisions of Infrastructures in Large-scale Geothermal Projects

The analysis in the previous sections reveals how the provision of infrastructures in their various forms differ in their nature, types, and quantity, depending on where they are provided, why they are provided, for whom they are provided, and who is providing them. Olkaria and Menengai geothermal projects are in semi-peripheral areas of Nakuru county where there were already some existing technical and social infrastructures (Fieldworks & interviews 2019, 2020).

In these areas, we see that more ‘generated’ infrastructures and relatively less ‘required’ infrastructures were provided. In contrast, in the case of Baringo-Silali project, which is in the peripheral and marginalized Baringo county, considerably more ‘required’ infrastructures had to be provided, as they were either too little or non-existent, in addition to the provided ‘generated’ infrastructures (Fieldworks & interviews 2019, 2020). So far, the total number of ‘required’ and ‘generated’ infrastructures as well as the capital and maintenance costs for providing them, are more for the Baringo-Silali geothermal projects in Baringo county when compared to Olkaria and Menengai geothermal projects in Nakuru county (Fieldworks & interviews 2019, 2020).

The provision of different ‘required’ infrastructures unveil interesting stakeholder involvement conditions and dynamics based on whose interest and purpose they serve. The reconstruction or tarmacking of already existing roads in semi-peripheral areas, which serve the interest of not only the project but also the interests of others actors, are often not solely delivered by the project developers and investors (GDC, KenGen, DFI interviews 2019). As the MSLR in the Olkaria geothermal projects illustrates, other actors or stakeholders who also benefit from the infrastructure make contributions for their construction and maintenance (Fieldwork 2019, Kuiper 2019). In contrast, the new roads usually constructed in formerly marginalized peripheries (example, the access roads for the Baringo-Silali project) as well as the water pump and storage systems provided in all the projects specifically serve the interest and purposes of the project and their developers at the time of their construction (GDC interviews, GDC 2019). As such, the project developers and investors in the geothermal project bore the sole responsibility of delivering the ‘required’ infrastructures.

Projects with more involvement of international development institutions and agencies as investors or financiers, so far, recorded a greater number of ‘generated’ infrastructures provision in form of corporate social responsibility and resettlement schemes for project affected persons (PAPs). There is currently more involvement of international development institutions and agencies in Olkaria and Menengai geothermal projects, and subsequently a greater number of ‘generated’ CSRs (Fieldwork 2019, 2020, GIBB Africa 2009, GDC 2010, 2013). However, this can be explained by the fact that Baringo-Silali project is just completing its exploratory stage. More CSR projects are expected to be provided in Baringo-Silali host communities in the future as the project proceeds into steam gathering and plant construction stages (GDC interviews 2020). By so doing, these development institutions and agencies (including the German KfW, the EIB, the AfDB and the French AFD) seek to establish their reputation as players who abides by sustainability principles (EIB, AfDB, KfW interviews 2019).

Furthermore, the level of engagements and negotiations between the project developers and host communities depends on whether the ‘required’ infrastructures are provided as a new project or as a reconstruction of already existing ones. The reconstruction of the MSLR roads leading to the Olkaria projects or the expansion of the roads leading to the Menengai Caldera, required less engagements with the host communities, except in specific cases where land had to be bought from their private owners (like in the case of Menengai geothermal projects) or in cases where project affected persons (PAPs) had to be resettled (like in the case of Olkaria geothermal projects) (Fieldwork 2019, 2020, GIB Africa 2009, GDC 2010, 2013). In contrast, the construction of new access roads for Baringo-Silali project development entailed constant and meticulous negotiations between the project developers and the host communities (Fieldwork 2019, GDC interviews 2019, Greiner et al. 2021). In this case, non-adherence to negotiated terms either due to change of contractors or ignorance of workers in the project sites often present protests and risks of conflicts (Fieldwork 2019, Klagge et al. 2020).

Conclusion

This paper analyzed the infrastructures in large-scale geothermal energy projects in Kenya, depicting their different processes and forms of materialization, and their complex socio-economic development linkages. We see how the materialization of 'core' infrastructure projects become generative, enabling the provision of other 'required' and 'generated' infrastructures. We also see that while the 'core' infrastructures of the projects are determined by and serve electrification interests at national level, their associated 'required' and 'dependent' infrastructures, mainly serve socio-economic development interests of project-host communities at local levels. Furthermore, by comparing the degree and scale of the provision of these infrastructures, the study reveals that the provision of these infrastructures is differentiated based on the local socio-economic and spatial contexts of the project-host communities. These findings demonstrate the complexity of sustainable large-scale projects planning and implementation in the Global South. It further shows how impact evaluation studies of large-scale development projects will be more encompassing and complete, when we consider the socio-spatial and socio-economic generativity potentials of their infrastructures. Overall, the socio-economic impacts of large-scale infrastructure projects are better appraised when the materialization and generativity potentials of the infrastructures are considered. The materialization of these infrastructures often leads to the emergence of other technical and social infrastructures – which also assume lives of their own, serving different interests. It is the combination of these infrastructures and their connections and interaction that allows for a more encompassed appraisal of the socio-economic impacts of large-scale infrastructure projects, especially in the global south context.

References

Addie J-PD, Glass MR Nelles J (2020) Regionalizing the infrastructure turn: a research agenda. *Regional Studies, Regional Science* 7(1): 10–26.

Aggarwal S (2018) Do rural roads create pathways out of poverty? Evidence from India. *Journal of Development Economics* 133(C): 375–395.

Albala-Bertrand JM, Mamatzakis EC (2004) The Impact of Public Infrastructure on the Productivity of the Chilean Economy. *Review of Development Economics* 8(2): 266–278.

Amin A (2006) The Good City. *Urban Studies* 43(5–6): 1009–1023.
Anand N (2011) Pressure: The politechnics of water supply in Mumbai. *Cultural Anthropology* 26(4): 542–564.
Anand N, Gupta A, Appel H (2018) *The Promise of Infrastructure*. Durham and London: Duke University Press.
Appiah DO, Sarfo M, Famieh B, Addai H (2017) Environmental and Socioeconomic Perturbations of a Dam Project on Catchment Communities, Ghana. *Global Environment Health and Safety* 1(2): 13.
Arrobbio A, Barros ACH, Beauchard RF, Berg AS, Brumby J, Fortin H, et al. (2014) *Corporate governance of state-owned enterprises: A toolkit (English).* Washington DC: World Bank Group.
Arrow KJ, Kurz M (1970) *Public Investment, the Rate of Return and Optimal Fiscal Policy*. Baltimore: The Johns Hopkins University Press.
Aschauer DA (1989). Is Public Expenditure Productive? *Journal of Monetary Economics* 23(2): 177–200.
Asher S, Novosad P (2020) Rural Roads and Local Economic Development. *American Economic Review* 110(3): 797–823.
Ballard R, Rubin M (2017) A marshall plan for human settlements: How megaprojects became South Africa's housing policy. *Transformation: Critical Perspectives on Southern Africa* 95(1): 1–31.
Banerjee S, Wodon Q, Diallo A, Pushak T, Uddin E, Tsimpo C, et al. (2008) *Access, affordability, and alternatives: Modern infrastructure services in Africa*. Washington DC: World Bank.
Barker A, Felliu E, Garcia-Blanco G, Kwiacinska K, Pedrola B (2021) Sustainability assessment of urban infrastructures. In E Crocci, B Lucchitta (eds.), *Nature Based Solutions for More Sustainable Cities: A Framework Approach for Planning and Evaluation*. Emerald Publishing Limited.
Barry A (2020) The Material Politics of Infrastructure. In S Maasen, S Dickel, C Schneider (eds.), *TechnoScienceSociety: Technological Reconfigurations of Science and Society, Sociology of Sciences Yearbook*, 91–109. Berlin: Springer.
Batey, P. W. J.; Madden, M. and Scholefield, G. 1993. Socioeconomic Impact Assessment of Large-scale Projects using Input–Output Analysis: A Case Study of an Airport. *Regional Studies* 27(3): 179–191.
Beevers L, Douven W, Lazuardi H, Verheij H (2012) Cumulative impacts of road developments in floodplains. *Transportation Research Part D: Transport and Environment* 17(5): 398–404.
Boyer D (2019) *Energopolitics*. Duke University Press.
Bryceson DF, Bradbury A, Bradbury T (2008) Roads to Poverty Reduction? Exploring Rural Roads' Impact on Mobility in Africa and Asia. *Development Policy Review* 26(4): 459–482.
Calderon C, Moral-Benito E, Servén L (2011) Is infrastructure capital productive? A dynamic heterogeneous approach. *Journal of Applied Econometrics* 30(2): 177–198.
Calderon C, Serven L (2008) *Infrastructure and Economic Development in Sub-Saharan Africa*. Policy Research Working Paper No. 4712. Washington, DC: World Bank.
Calderon C, Chong A (2009) Labor Market Institutions and Income Inequality: An Empirical Exploration. *Public Choice* 138(1): 65–81.
Campbell M, Alamgir M, Laurance W (2017) Roads to ruin. *Australasian Science* 38: 40–41.
Chambers J (2019) Fluidities and Fixities: Examining the Alignment of Digital Platforms Within Nairobi's Heterogeneous Infrastructural Configurations. *Front. Sustain. Cities* 1: 8.

Chattopadhyay S (2012) *Unlearning the City. Infrastructure in a New Optical Field*. Minneapolis, MN: University of Minnesota Press.

Cirolia LR (2020) Fractured fiscal authority and fragmented infrastructures: Financing sustainable urban development in Sub-Saharan Africa. *Habitat International* 104: 102233.

Cirolia LR, Smit W (2017) Fractured approaches to urban transformation: Analyzing parallel perspectives in South Africa. *Transformation: Critical Perspectives on Southern Africa* 95(1): 63–84.

Coffin AW (2007) From roadkill to road ecology: a review of the ecological effects of roads. *J. Transp. Geography* 15: 396–406.

Collier E, Maussion F, Nicholson LI, Mölg T, Immerzeel WW, Bush ABG (2015) Impact of debris cover on glacier ablation and atmosphere–glacier feedbacks in the Karakoram. *The Cryosphere* 9: 1617–1632.

Cooper F (1996) *Decolonization and African society: The labor question in French and British Africa* (Vol. 89). Cambridge, MA: Cambridge University Press.

Coutard O, Rutherford J (2011) The rise of post-networked cities in Europe? Recombining infrastructural, ecological and urban transformations in low carbon transitions. In H Bulkeley, V Castán Broto, M Hodson, S Marvin (eds.), *Cities and Low Carbon Transitions*, 104–123. London: Routledge.

Coutard O, Rutherford J (Eds.) (2015) *Beyond the Networked City: Infrastructure Reconfigurations and Urban Change in the North and South*. Abingdon: Routledge.

De Boeck F, Baloji S (2016) *Suturing the city: Living together in Congo's urban worlds*. London: Autograph ABP.

Dethier JJ, Hirn M, Straub S (2008) *Explaining Enterprise Performance in Developing Countries with Business Climate Survey Data*. Policy Research Working Paper, 4792. Washington, DC: World Bank.

Divine OA, Michael S, Bridgette F, Henrietta A (2017) Environmental and Socioeconomic Perturbations of a Dam Project on Catchment Communities, Ghana. *Global Environment Health and Safety* 1(2): 13.

Easterly W, Rebelo S (1993) Fiscal Policy and Economic Growth: An Empirical Investigation. *Journal of Monetary Economics* 32: 417–458.

Edwards PN (2002) Infrastructure and Modernity: Scales of Force, Time, and Social Organization in the History of Sociotechnical Systems. In TJ Misa, P Brey, A Feenberg (eds.), *Modernity and Technology*, 185–225. Cambridge, MA: MIT Press.

Enns C, Bersaglio B (2020) On the Coloniality of 'New' Mega-Infrastructure Projects in East Africa. *Antipode* 52(1): 101–123.

Ernstson H, Lawhon M, Duminy J (2014) Conceptual vectors of African urbanism: 'Engaged theory-making' and 'platforms of engagement'. *Regional Studies* 48(9): 1563–1577.

Estache A, Garsous G (2012) *The impact of infrastructure on growth in developing countries*. Waghington DC: IFC Economics Notes.

Estache A, Fay M (2010) Current Debates on Infrastructure Policy. In M Spence, D Leipziger (eds.), *Globalization and Growth, Implications for a Post-Crisis World*, 151–193. Washington DC: Commission on Growth and Development.

Estache A, Wren-Lewis L (2009) Toward a Theory of Regulation for Developing Countries: Following Jean-Jacques Laffont's Lead. *Journal of Economic Literature* 47(3): 730–771.

Foley JA, Asner GP, Costa MH, Coe MT, DeFries R, Gibbs HK, et al. (2007) Amazonia revealed: forest degradation and loss of ecosystem goods and services in the Amazon Basin. *Frontiers in Ecology and the Environment* 5: 25–32.

Foster V, Briceno-Garmendia C (Eds.) (2010) *Africa's infrastructure: A time for transformation. Africa development forum*. Washington DC: World Bank.

Furlong K (2020) Geographies of infrastructure 1: Economies. *Progress in Human Geography* 44(3): 572–582.

Garsous G (2012) *How Productive is Infrastructure? A Quantitative Survey*. ECARES Working Paper. Université libre de Bruxelles.

GDC (2010) *Kenya Goes Big on Geothermal Energy: thrill as GDC begins to drill*. Steam magazine Issue No. 1, Jan-March 2010.

GDC (2013) *Charm of Menengai: Governor, MPs, senators enchanted*. Steam magazine Issue No. 9, Sept-Nov 2013.

GDC (2019) *The Baringo Bloom Powering the North: GDC transforming Baringo*. Steam magazine Issue No. 14, March 2019.

GIBB Africa (2009) *KENGEN Olkaria IV Project - Resettlement Action Plan*. Nairobi, Kenya.

Glass MR, Addie J-PD, Nelles J (2019) Regional infrastructures, infrastructural regionalism. *Regional Studies* 53(12): 1651–1656.

Goodfellow T (2020) Finance, infrastructure, and urban capital: The political economy of African 'gap-filling'. *Review of African Political Economy* 47(164): 256–274.

Graham S, McFarlane C (Eds.) (2014) *Infrastructural Lives: Urban Infrastructure in Context*. Abingdon: Routledge.

Greiner C (2016) Land-use change, territorial restructuring, and economies of anticipation in dryland Kenya. *Journal of Eastern African Studies* 10(3): 530–547.

Greiner C, Greven D, Klagge B (2021) Roads to change: Livelihoods, land disputes, and anticipation of future developments in rural Kenya. *European Journal of Development Research* 33: 1044–1068.

Hall JW, Henriques JJ, Hickford AJ, Nicholls RJ (2013) Systems-of-systems analysis of national infrastructure. *Proc. ICE Eng. Sustain*. 166(5): 249–257.

Howe C (2019) *Ecologics*. Duke University Press.

Howe C, Lockrem J, Appel H, Hackett E, Boyer D, Hall R, et al. (2015) Paradoxical Infrastructures: Ruins, Retrofit, and Risk. *Science, Technology, & Human Values* 41(3): 1–19.

Jacoby HG (2000) Access to Markets and the Benefits of Rural Roads. *The Economic Journal* 110(465): 713–737.

Jaglin S (2015) Is the network challenged by the pragmatic turn in African cities? In O Coutard, J Rutherford (eds.), *Beyond the Networked City: Infrastructure Reconfigurations and Urban Change in the North and South*, 182–203. Abingdon: Routledge.

Kanai JM, Schindler S (2019) Peri-urban promises of connectivity: Linking project-led polycentrism to the infrastructure scramble. *Environment and Planning A: Economy and Space* 51(2): 302–322.

Kenley R, Harfield T, Bedggood J (2014) Road Asset Management: The role of location in mitigating extreme flood maintenance. *Procedia Economics and Finance* 18: 198–205.

Kimari W, Ernstson H (2020) Imperial remains and imperial invitations: Centering race within the contemporary large-scale infrastructures of east Africa. *Antipode* 52(3): 825–846.

Klagge B, Nweke-Eze C (2020) Financing large-scale renewable-energy projects in Kenya: investor types, international connections, and financialization. *Geografiska Annaler: Series B, Human Geography* 102(1): 61–83.

Klagge B, Greiner C, Greven D, Nweke-Eze C (2020) Cross-scale Linkages of Centralized Electricity Generation: Geothermal Development and Investor-community Relations in Kenya's Semi-arid North. *Politics and Governance* 8(3): 211–222.

Korytárováa J, Hromádkaa V (2014) The Economic Evaluation of Megaprojects – Social and Economic Impacts. In *Procedia - Social and Behavioral Sciences* 119: 495–502.

Kuiper G (2019) *Agro-Industrial Labour in Kenya: Cut Flower Farms and Migrant Workers' Settlements*. Cham: Palgrave Macmillan.

Larkin, B (2008) Infrastructure, the Colonial Sublime, and Indirect Rule. *Signal and Noise: Media, Infrastructure, and Urban Culture in Nigeria*, 16-47. New York, USA: Duke University Press.

Larkin B (2013) The Politics and Poetics of Infrastructure. *Annual Review of Anthropology* 42: 327-343

Laurance WF, Croes BM, Tchignoumba L, Lahm SA, Alonso A, Lee ME, et al. (2006) Impacts of roads and hunting on Central African rainforest mammals. *Conserv. Biol*. 20: 1251–1261.

Lawhon M, Ernstson H, Silver J (2014) Provincializing urban political ecology: Towards a situated UPE through African urbanism. *Antipode* 46(2): 497–516

Lawhon M, Nilsson D, Silver J, Ernstson H, Lwasa S (2018) Thinking through heterogeneous infrastructure configurations. *Urban Studies* 55(4): 720–732.

Lindell I (2008) The multiple sites of urban governance: Insights from an African city. *Urban Studies* 45(9): 1879–1901.

Luiz J (2010) Infrastructure investment and its performance in Africa over the course of the twentieth century. *International Journal of Social Economics* 37(7): 512–536.

Mandle L, Tallis H, Sotomayor L, Vogl AL (2015) Who loses? Tracking ecosystem service redistribution from road development and mitigation in the Peruvian Amazon. *Frontiers in Ecology and the Environment* 13: 309–315.

Maringanti A, Jonnalagadda I (2015) Rent gap, fluid infrastructure and population excess in a gentrifying neighbourhood. *City* 19: 365–374.

McFarlane C, Silver J (2017) Navigating the city: Dialectics of everyday urbanism. *Transactions of the Institute of British Geographers* 42(3): 458–471.

Meehan KM (2014) Tool-power: Water infrastructure as wellsprings of state power. *Geoforum* 57: 215–224.

Monstadt J, Schramm S (2017) Toward the Networked City? Translating Technological ideals and Planning Models in Water and Sanitation Systems in Dar es Salaam. International Journal of Urban and Regional Research 41(1): 104–125.

Mu R, van de Walle D (2011) Rural Roads and Local Market Development in Vietnam. *Journal of Development Studies* 47(5): 709–734.

Nweke-Eze C, Adongo C (forthcoming) Geothermal development in Olkaria: Between national development visions and local community socio-economic realities. In Bollig et al. (ed.), *Agricultural intensification and resource conflicts: Social dynamics in the Lake Naivasha Basin*. James Curry.

Nweke-Eze C, Kioko EM (2021) *'But we cannot do it all': Investors' sustainability tensions and strategic selectivity in the development of geothermal energy in Kenya*. In W Leal Filho et al. (eds.), *Sustainable Development in Africa*, 385–404. World Sustainability Series. Springer, Cham.

Ogola PFA (2013) *The power to change: Creating lifeline and mitigation adaptation opportunities through geothermal energy utilization*. PhD Dissertation. University of Iceland.

Pieterse E (2019) *The potential for sustainable urbanisation in Africa*. Alfred Herrhausen Gesellschaft.

Schade J (2017) *Kenya 'Olkaria IV' Case Study Report: Human Rights Analysis of the Resettlement Process*. Working Paper 151. Bielefeld: Centre on Migration, Citizenship and Development (COMCAD).

Schouten MAC, Mathenge RW (2010) Communal sanitation alternatives for slums: A case study of Kibera, Kenya. *Physics and Chemistry of the Earth, Parts A/B/C* 35(13): 815–822.

Scott JC (1998) *Seeing Like a State: How Certain Schemes to Improve the Human Condition have Failed*. New Haven, CT: Yale University Press.

Seethepalli K, Bramat MC, Veredas D (2008) *How Relevant Is Infrastructure to Growth in East Asia?* The World Bank Policy Research Working Paper 4597.

Silver J (2014) Incremental infrastructures: Material improvisation and social collaboration across post-colonial Accra. *Urban Geography* 35(6): 788–804

Silver J, Marvin S (2017) Powering sub-Saharan Africa's urban revolution: An energy transitions approach. *Urban Studies* 54(4): 847–861.

Simone A (2004) People as infrastructure: Intersecting fragments in Johannesburg. *Public Culture* 16(3): 407–429.

Star SL 1999. The Ethnography of Infrastructure. *American Behavioral Scientist* 43(3): 377–391.

Straub S (2008) *Infrastructure and Growth in Developing Countries: Recent Advances and Research Challenges*. World Bank Policy Research Working Paper Series. No. 4460. Washington, DC: World Bank.

Straub S (2011) Infrastructure and Development: A Critical Appraisal of the Macro-level Literature. *The Journal of Development Studies* 47 (5): 683–708.

Straub S, Terada-Hagiwara A (2011) Infrastructure and Growth in Developing Asia. *Asian Development Review* 28(1): 119–156.

Talen E (2019) Neighbourhoods. Oxford: Oxford University Press.

Terrefe B (2020) Urban layers of political rupture: The 'new' politics of Addis Ababa's megaprojects. *Journal of Eastern African Studies*.

Thekdi SA, Chatterjee S (2019) Toward adaptive decision support for assessing infrastructure system resilience using hidden performance measures. *J. Risk Res.* 22: 1020–1043.

Trombulak SC, Frissell CA (2001) Review of Ecological Effects of Roads on Terrestrial and Aquatic Communities. *Conservation Biology* 14 (1): 18-30.

Turner C (2018) *Regional Infrastructure Systems: The Political Economy of Regional Infrastructure*. Edward Elgar Publishing Limited.

Van Noorloos F, Kloosterboer M (2018) Africa's new cities: The contested future of urbanisation. *Urban Studies* 55(6): 1223–1241.

Weinhold D, Reis E (2008) Transportation costs and the spatial distribution of land use in the Brazilian Amazon. *Global Environmental Change* 18(1): 54–68.

World Bank (1994) *Infrastructure for Development*. Washington, DC: World Development Report.