

Machine learning-based models for forecasting radio refractivity over the coastal area of South Africa

Yusuf Babatunde Lawal^{1*}, Pius Adewale Owolawi², Chunling Tu³, Etienne Van Wyk⁴, Joseph Sunday Ojo⁵

^{1,2,3}Department of Computer Systems Engineering, Tshwane University of Technology, South Africa; lawalyb@tut.ac.za (Y. B. L.), owolawipa@tut.ac.za (P. A. O.), duc@tut.ac.za (C. T.).

⁴Faculty of Information and Communications Technology, Tshwane University of Technology, South Africa; vanwykea@tut.ac.za (E. V. W.).

⁵Department of Physics, Federal University of Technology, Akure, Nigeria; ojojs_74@futa.edu.ng (J. S. O.).

Abstract: Surface refractivity is a crucial parameter that determines the bending of radio signals as they propagate within the troposphere. It is greatly influenced by the atmospheric weather conditions and changes rapidly, especially in the coastal areas. This research utilized 50 years (1974–2023) surface temperature, pressure, and humidity data from six coastal stations in South Africa to forecast radio refractivity in the Mediterranean climate. Five machine learning models: Gated Recurrent Unit (GRU), Light Gradient Boosting Machine (LightGBM), Long-Short Term Memory (LSTM), Prophet, and Random Forest were trained for future prediction of surface refractivity at any coastal area in South Africa. The stations latitude, longitude, altitude, surface refractivity and date were applied as the input parameters to train the models. The models were optimized through the randomized searchCV hyperparameter tuning to improve their efficiency. The LightGBM outperformed other models with RMSE and adjusted determination coefficients of 1.67 and 0.96, respectively. The model is recommended for future prediction of surface refractivity needed for the improvement of point-to-point wireless communication, terrestrial radio and television transmissions, and mobile communication networks in the coastal sub-tropical regions.

Keywords: GRU, Hyperparameter tuning, Light GBM, LSTM, Machine learning, Surface refractivity, Troposphere.

1. Introduction

Surface refractivity is crucial for accurate modeling of radio wave propagation, particularly in applications such as telecommunications and meteorology [1]. Accurate refractivity gradient data is essential for radio system design, as strong negative gradients can cause interference, while positive gradients affect antenna height requirements [2]. Understanding the variations in refractivity due to factors like humidity, temperature, and atmospheric pressure is essential for improving signal strength and clarity. These variations can be influenced by geographical features, seasonal changes, and even time of day, making it imperative to gather comprehensive data for precise estimations [3]. In coastal regions, the proximity to large bodies of water adds another layer of complexity, as sea breezes, ocean currents and evaporation can significantly alter local atmospheric conditions. This necessitates the application of advanced modeling techniques to forecast local radio refractivity, ensuring accurate and reliable propagation data for optimization of terrestrial communication links.

Several studies have shown that temperature, humidity, and height above ground level significantly influence refractivity, with an inverse relationship observed between refractivity and signal strength [4]. The existing refractive index model is deemed inadequate for Southern Africa, necessitating improved propagation models for better spectrum usage, particularly in mobile systems [5]. Accurate refractivity gradient data is essential for radio system design, as strong negative gradients can cause

interference, while positive gradients affect antenna height requirements [2]. The current research aims to enhance understanding of refractivity profiles and their impact on radio communications.

Most previous studies have recommended fixed mean values through some statistical and advanced prediction techniques. For instance, Dabideen et al [6] derived surface refractivity and 1km-height refractivity gradient from Kwazulu-Natal atmospheric data obtained from the South African Weather Station (SAWS). The research recommends monthly mean refractivity in the range of 320–360 N-units. The 50 years atmospheric data used for the current research has revealed that surface refractivity in any of the study stations vary beyond this range especially during extreme atmospheric conditions.

Asiyo and Afullo [7] computed refractivity and k-factor using three years radiosonde data from five stations across South Africa. The results were applied through the interpolation techniques to develop and recommend k-factor contour map of South Africa. Seasonal variability of atmospheric parameters such as temperature, pressure, and humidity could influence great difference between the computed and real-time values of surface refractivity which is essential for improved dynamic communication systems. Hence, the need for a dynamic prediction method.

Salamon et al [2] applied a new empirical model for the cumulative distribution of meteorological data to predict surface refractivity and refractivity gradient in the lowest 80 m of the atmosphere. Sixteen years (1997–2012) radiosonde data from some stations excluding South Africa was obtained from the NOAA/ESRL Radiosonde Database. Although, the research recommended a new model for surface refractivity prediction, spatial variation of atmospheric parameters makes the model unfit for the mediterranean climate in South Africa. Agbo [8] applied Multiple linear regression model to predict refractivity in the Calabar. The derived partial differential equations require temperature, pressure and humidity as input to compute predicted refractivity. The model performed well with minimal error but requires the measured values of the three atmospheric parameters for surface prediction. Moreover, the model which was trained with subtropical region data from Calabar (Nigeria) is not suitable for the newly selected research locations.

This research presents a new approach for forecasting surface refractivity in the coastal regions of South Africa.

1.1. Overview of the Machine Learning Models

1.1.1. The Gated Recurrent Unit Model

It is a kind of Recurrent Neural Network RNN that can effectively solve long-term dependencies issues and does not pose a vanishing gradient problem that is associated with the traditional RNN. The GRU improves the RNN functionality by adopting the use of gating mechanism which controls information flow, making the network capable of holding essential information over time and offering efficient training [9]. The GRU is made up of four major components: the reset gate, the update gate, the current memory content and the final hidden state. The two gates are responsible for information storage management by ensuring secured storage and prompt update. The current memory content is the candidate hidden state which combines its content with that of the reset gate to arrive at the final hidden state. The final hidden state is derived through feature interpolation between the current memory value and previous hidden state. Its operation is controlled by the update gate thus, they can lower the data quantity with improved accuracy. The final hidden state G_i of the GRU is computed using equation (1).

$$G_i = (1 - x_i) \odot G_{i-1} + x_i \odot \tilde{G}_i \quad (1)$$

where x_i is the update gate, \odot is the element-wise multiplication, G_{i-1} is the previous hidden state, \tilde{G}_i is the current memory content [9]. GRU algorithm is less complex compared to LSTM but provides efficient computations for temporal sequence analysis.

1.1.2. The Light Gradient Boosting Machine (LightGBM)

It is an advanced extensive gradient boosting for machine learning classification, regression, ranking, and so on, which was developed by Microsoft. LightGBM is particularly fast and efficient, especially when enormous data and multi features are involved. It does this through changes to the ways trees are built, such as the introduction of histogram-based decision tree learning and the complete family of leaf-wise trees for speedier and more accurate training. Elimination of the continuous need to process raw values directly allows for a decrease in memory footprint as well as faster training. This is done using the histogram-based technique whereby quantitative data values are grouped in bins. The leaf-wise feature of the LightGBM helps to focus only on leaf with highest probability of loss reduction thereby reducing the number of splits and improving accuracy. The objective function F_{obj} of the LightGBM is minimized at each stage of iteration by equation (2) while equation (3) gives the total output prediction $\hat{x}_i^{(t)}$ [10] & [11].

$$F_{obj} = \sum_{i=1}^n l(x_i, \hat{x}_i) + \sum_{j=1}^m \Omega(f_j) \quad (2)$$

$$\hat{x}_i^{(t)} = \hat{x}_i^{(0)} + \eta \sum_{m=1}^t g_m(h_i) \quad (3)$$

where $l(x_i, \hat{x}_i)$ is the loss function, $\Omega(f_j)$ is the regularization term which prevents overfitting, \hat{x}_i is the predicted output, $\hat{x}_i^{(0)}$ is the baseline prediction, η is the contribution control which lies between 0 and 1, g_m is the tree function for iteration m , h_i is the input value.

1.1.3. The Long Short-Term Memory (LSTM) Model

The LSTM model is a kind of Recurrent Neural Network (RNN) commonly used for learning comparatively long-term dependencies based on previous data trend. Unlike the traditional RNNs, LSTMs are equipped with special units called memory cells. Information is stored in the memory cell for a long period. Each cell has three main gates: the input gate, which decides what information is allowed into updating the cell state. The forget gate identifies the remaining part of the previous data saved in the memory cell and ensure it is cleared before the next iteration. The output gate manages the information flow rate from the memory cell that is used to compute the output. The activation vector used for the computation of the output gate is presented in equation (4) [12].

$$o_t = \sigma(W_o \cdot [g_{t-1}, y_t] + c_o) \quad (4)$$

where W_o is the output gate weight matrix, g_{t-1} is the previous hidden state, y_t is the input at the current time step, c_o is the bias term for the candidate memory cell state.

1.1.4. The Prophet Model

Prophet is an automated forecasting tool developed by Facebook researchers, Sean J. Taylor and Ben Letham in 2017 [13]. It is suitable for time series data with daily observations and noticeable seasons. The model is resistant to missing observations problems and shifts in trend as it handles outliers and seasonality appropriately. Being the most appropriately used in the analysis of business time series data, the model can be applied in sales, marketing, and web traffic data. Prophet uses additive approach to model times-series function as a combination of data trend, seasonality, and holiday effects. The additive model equation $F(t)$ is defined by equation (5).

$$F(t) = s(t) + u(t) + h(t) + \epsilon_t \quad (5)$$

where $s(t)$ is the seasonality function that captures periodic pattern of the data, $u(t)$ is the trend function which models the non-periodic changes in the time series, $h(t)$ accounts for the holidays such as special events or anomalies in the time series, ϵ_t is the error term responsible for unexplained variability [13, 14]. The seasonality function as shown in equation (6), uses the Fourier series which enables the model to capture repetitive patterns like daily, weekly, monthly, bi-annual, annual, variations.

$$s(t) = \sum_{i=1}^n \left[a_i \cos\left(\frac{2\pi it}{T}\right) + b_i \sin\left(\frac{2\pi it}{T}\right) \right] \quad (6)$$

where T is period for the seasonality, n is the number of terms in the fourier series, a_i and b_i are the coefficients of the fourier series learned from the data. Prophet uses two types of trend model $u(t)$ to learn data trend pattern; the piecewise linear trend and logistic growth trend models expressed in equations (7) and (8) respectively.

$$u(t) = (k + a(t)^T \delta)t + (m + a(t)^T \gamma) \quad (7)$$

where k is the initial growth rate, $a(t)$ is the indicator function that determines when the change occurs, δ is the change in growth rate at change points, m is intercept of the linear function (i.e offset parameter), γ is the adjustment vector which counterbalances the changes in the trend.

$$u(t) = \frac{C}{1 + \exp(-k(t-m))} \quad (8)$$

where C is the carrying capacity or the maximum possible value, k is the initial growth rate, m is the time at which the growth rate reaches maximum or midpoint.

The piecewise linear trend is a linear function with potential at a point where the trend's rate of growth changes. The logistic growth is commonly used to activate rapid growth and saturate the model at a maximum capacity. The holiday effects which have significant influence on time series is computed using:

$$h(t) = \sum_{i=1}^n \alpha_i D_i(t) \quad (9)$$

where α_i is the impact of each holiday, $D_i(t)$ is an indicator function which is 1 if day t is holiday and 0 if it is non-holiday? Prophet usually assumes the error term ϵ_t to follow a normal distribution.

1.1.5 The Random Forest Model

The Random Forest Model is a machine learning algorithm that randomizes the use of a disentangled set of several decision trees to arrive at a special product. It is a type of learning model that is relatively appropriate for classification and regression analysis. Its working principle depends on establishing several decision trees during the training phase. After the training, it comes up with its decisions which are the class modes for classification analysis or the mean of the individual trees' prediction for regression analysis using equations (10) and (11) respectively [15].

$$\hat{G}_i = \text{mode}\{T_i^b(P)\}_{b=1}^B \quad (10)$$

$$\hat{G}_i = \frac{1}{B} \sum_{b=1}^B T_i^b(P) \quad (11)$$

where \hat{G}_i is the final prediction for input sample P . B is the total trees in the forest. T_i^b is the prediction from the b^{th} tree for the i^{th} sample P . The model takes advantage of the multiple decision trees in the forest to improve prediction accuracy.

2. Methodology

The surface refractivity N was computed from the surface temperature, pressure, and humidity retrieved from Era5 dataset of the ECMWF using equation (12) [16].

$$N = \frac{77.6P}{T} + \frac{3.73 \times 10^5 e}{T^2} \quad (12)$$

where P is the total atmospheric pressure in hPa, T is the air temperature in Kelvin and e is the water vapour pressure in hPa defined by equation (13).

$$e = H \times \frac{6.1121 \exp\left(\frac{17.502t}{t+240.97}\right)}{100} \quad (13)$$

where t is the air temperature in in degrees Celsius.

The seasonal trend of the computed surface refractivity was studied and discussed based on the four prevailing seasons in South Africa [17] & [18]. The coordinates of the study stations presented in Table 1 were obtained from altitude-maps database [19]. The stations' latitude, longitude, altitude and date are the independent variables which serve as the input variables for the model training. The surface refractivity which is the dependent variable was fed into the models using python programming language. The data was split into 80% for training and 20% for testing. The initial results were optimized using the randomized searchCV hyperparameter tuning. The models performance metrics such as RSME and determination co-efficient were compared to identify the best model. The future prediction python script which any location coordinates and future date as input parameters was developed for forecasting corresponding surface refractivity.

3. Results and Discussion

3.1. Seasonal variation of Surface refractivity

Selected time series plots of the 50 years surface refractivity data used for training the models are presented in Figures 1 and 2. Durban and Cape Town are the selected typical stations whose monthly refractivity values are shown in Figures 1 and 2, respectively. Comparison of the stations' seasonal variation could not be revealed in the figures due to the crowded data points for the 50 years. To visualize the seasonal trend of the data, only five years data were plotted for all the stations and presented in Figures 3.

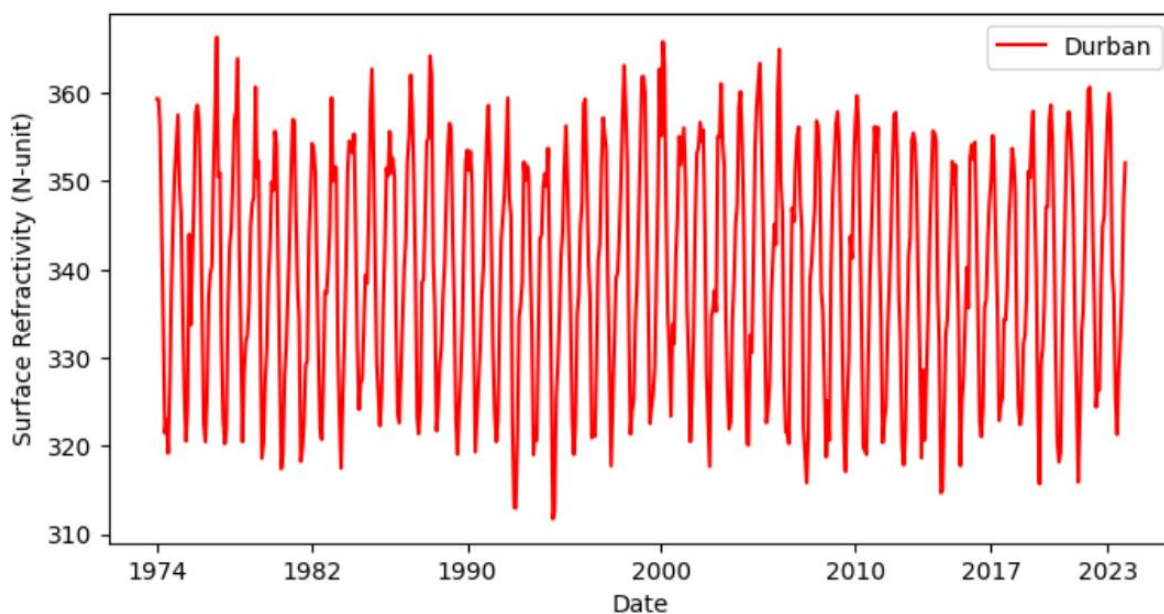


Figure 1. The time series plot of the 50 years surface refractivity data at Durban used for training the models.

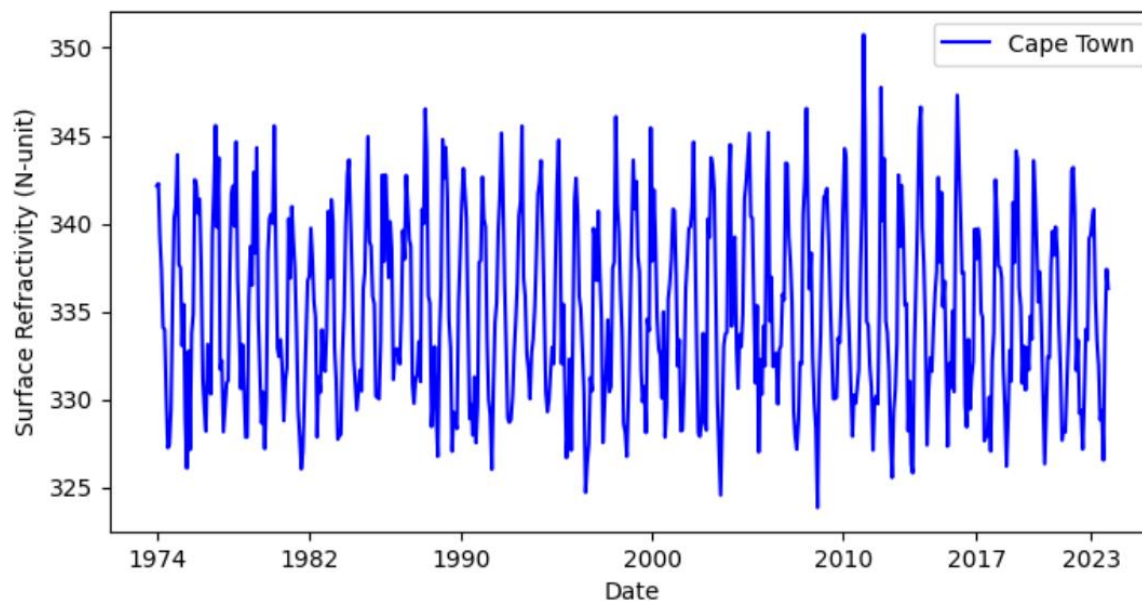


Figure 2.
The time series plot of the 50 years surface refractivity data at Cape Town used for training the models.

The 5 years' time series plot presented in Figures 3 shows that surface refractivity varies significantly over the four seasons in South Africa annually. Surface refractivity across all the stations usually attains all-time lowest during the winter season (June-August). This is because relative humidity reaches maximum due to high rate of evaporation, frequent heavy rainfalls, and cool temperature [20] & [21]. The East London has the lowest mean value of about 310 N-units, followed by 316 N-units at Port Elizabeth, 322 N-units at Durban and Mossel Bay, 330 N-units at Cape Town while Richard Bay recorded highest value of 332 N-units. The spring season (September-November) witnessed gradual reduction in rainfall as the coastal temperature rises due to transition from winter to summer. The coastal region, especially the southeastern part experiences steady rise in moisture from the Agulhas current, with humidity often reaching about 70% by late spring [22]. The water vapour pressure which has a significant influence on radio refractivity becomes high [23]. Consequently, surface refractivity rises gradually which signifies the on-set of the summer season. Surface refractivity reaches its peak in the summer (December-February) as the hot dry air dominates the coastal region. The average peak values during the study period are 361, 358, 353, 345, 341, and 338 N-units at Richard Bay, Durban, East London, Port Elizabeth, Mossel Bay, and Cape Town respectively. Refractivity begins to decline during Autumn (March-May) due to gradual cooling temperature and reduced relative humidity. The observed seasonal trend and values resonate with the results of Asiyu [24] that utilized ten years (2001-2010) radiosonde data for Cape Town, Durban and Port Elizabeth.

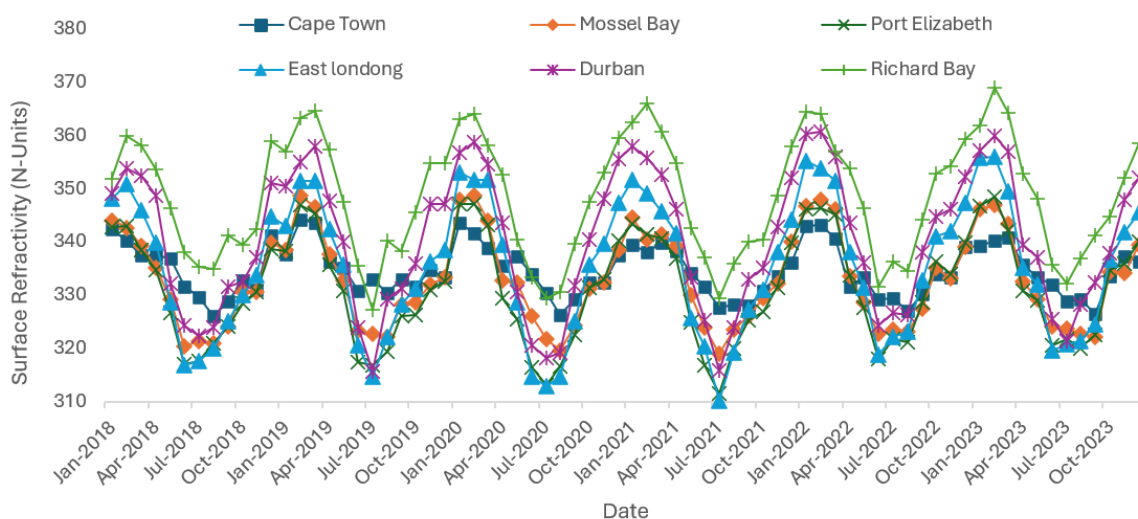


Figure 3.
Five-year time-series plot of surface refractivity of the sample data.

3.2. Analysis of the Prediction Models

Table 2 presents the performance metrics of the five models employed for the prediction of surface refractivity. The model performance metrics show notable differences in terms of accuracy for forecasting surface refractivity. The LightGBM possesses the best performance metrics with the lowest MAE and MSE and highest adjusted R^2 of 0.96, followed by Random Forest as shown in Table 2. This indicates that LightGBM captures the variation patterns in the data very effectively, explaining approximately 97% of the variance.

Table 2.
Performance metrics of the trained models.

Model	MAE	RMSE	Adjusted R^2
Random forest	2.69	5.62	0.92
LightGBM	1.67	3.84	0.97
LSTM	3.14	10.33	0.88
GRU	3.13	10.56	0.89
Prophet	14.46	14.86	-1.56

It can be inferred that the gradient boosting feature of the LightGBM allows it to handle seasonal variation and complex relationships in the data efficiently, making it an excellent choice for this task. Random Forest also performs well, but not as well as LightGBM. It has slightly higher errors and a lower Adjusted R^2 of 0.92, indicating that it explains around 92% of the variance. While the Random Forest is quite accurate, its ensemble of decision trees may not capture subtler patterns in the data as effectively as LightGBM. The LSTM model's performance is decent, with an Adjusted R^2 of 0.88. This indicates it explains 88% of the variance. However, its error rates are higher than both LightGBM and Random Forest. LSTM is well-suited for sequential data, but in this case, it may not have captured the seasonal patterns as effectively, or it could require further tuning. The GRU model performs similarly to LSTM, with nearly identical error metrics and an Adjusted R^2 of 0.89. This model captures around 89% of the variance but still lags behind LightGBM and Random Forest. GRU, being simpler than LSTM, is generally faster but may not provide additional benefits in this specific context. Prophet performs very poorly in this case, with high error values and a negative Adjusted R^2 , indicating that it fails to explain the variance in the data. The negative Adjusted R^2 implies that the model performs

worse than a simple mean model. Hence, the model is unsuitable for this complex multivariate dataset modeling, as it fails to capture the data's patterns effectively.

The recommended LightGBM model was used to forecast refractivity for the remaining months in 2024 and 2025 in several locations within the coastal stations. The results obtained and the percentage errors for a typical (Cape Town) are presented in Figure 4 and Table 2. The figure clearly illustrates that the model provides an efficient prediction method. The maximum percentage error between the measured and predicted across all tested stations within the coastal region is 1.2%.

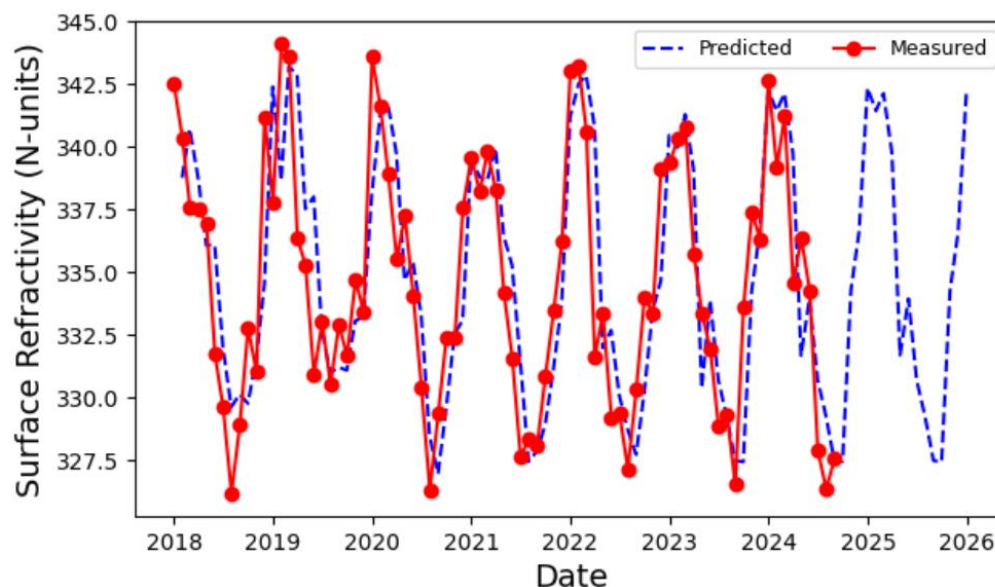


Figure 4. Time-series of the LightGBM trained model showing the measured and predicted surface refractivity for a typical station; Cape Town.

4. Conclusion

The preliminary stage of this research studies the seasonal variation of surface refractivity over the coastal region of South Africa. The observed seasonal trend is in consonant with previous research works. Subsequently, fifty years atmospheric data of six coastal locations in South Africa were used to train five machine learning models for surface refractivity forecasting. Comparative analysis of the model performance recommends the LightGBM model with maximum percentage error of about 1.2% based on tested future predictions. The model could be used to derive past, real-time and future values of surface refractivity in areas where there is scarcity or absence of atmospheric data. The models could be applied to other climatic regions such as arid, semi-arid, tropical climate to ascertain its suitability.

Funding:

This research is supported by Tshwane University of Technology South Africa.

Copyright:

© 2024 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

References

- [1] I. Alam, M. Najam-ul-Islam, U. Mujahid, S. A. A. Shah, and R. U. Haq, "Refractivity variations and propagation at Ultra High Frequency," *Results in Physics*, 2017. doi: 10.1016/J.RINP.2017.09.050.
- [2] S. J. Salamon, H. J. Hansen, and D. Abbott, "Prediction of surface refractivity gradient distributions, from weather station surface data," in the 8th European Conference on Antennas and Propagation (EuCAP 2014), The Hague, Netherlands, Apr. 2014, pp. 296–300. doi: 10.1109/eucap.2014.6901750.
- [3] Y. B. Lawal and E. T. Omotoso, "Investigation of Point Refractivity Gradient and Geoclimatic Factor at 70 m Altitude in Yenagoa, Nigeria," *Journal of the Nigerian Society of Physical Sciences*, vol. 5, no. 1, p. 1081, 2023. doi: 10.46481/jnsps.2023.1081.
- [4] S. S. Hussain, L. Ali, A. Hussain, A. Ullah, L. Sun, and G. Zhang, "Surface refractivity profile construction on one-fourth kilometer square area for 1800 to 1900 MHz frequency," *IEEE Access*, vol. 8, pp. 35650–35661, 2020. doi: 10.1109/ACCESS.2020.2973241.
- [5] A. J. Palmer and D. C. Bake, "Towards a model of the radio refractive index for Southern Africa-current status and future perspectives," in 1999 IEEE Africon. 5th Africon Conference in Africa, Cape Town, South Africa, Sep. 1999. doi: 10.1109/AFRCON.1999.821921.
- [6] A. S. Dabideen, M. Gopichund, and T. J. Afullo, "Radio refractivity distribution and duct and fading occurrence measurements in Kwazulu-Natal," *SAIEE Africa Research Journal*, vol. 96, no. 2, pp. 121–133, Jun. 2005. doi: 10.23919/SAIEE.2005.9488095.
- [7] M. O. Asiyo and T. J. O. Afullo, "Spatial interpolation for mapping geoclimatic factor K in South Africa," in *PIERS Proceedings*, Taipei, Taiwan, Mar. 2013, pp. 25–28.
- [8] P. E. Agbo, "The role of statistical methods and tools for weather forecasting and modeling," *IntechOpen*, 2021. doi: 10.5772/intechopen.96854.
- [9] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.
- [10] X. Zhang, L. Zhang, D. Zhao, and Y. Zhang, "Efficient parallel gradient boosting decision trees," *IEEE Transactions on Big Data*, vol. 7, no. 2, pp. 234–246, Jun. 2021. doi: 10.1109/TBDDATA.2018.2884052.
- [11] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, pp. 3149–3157.
- [12] C. Aggarwal, *Neural Networks and Deep Learning*, 1st ed. Springer Link, 2023, pp. 265–304. doi: 10.1007/978-3-031-29642-0_8.
- [13] S. J. Taylor and B. Letham, "Prophet: Rapid automatic forecasting of time series data," in *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.
- [14] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018. doi: 10.1080/00031305.2017.1380080.
- [15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.
- [16] ITU-R, "The radio refractivity index: Its formula and refractivity data," *Institution of Telecommunication Union*, Geneva, Switzerland, Recommendation of ITU-R P. 453-9, Feb. 2012.
- [17] M. R. Jury, "Climate trends in southern Africa," *South African Journal of Science*, vol. 110, no. 1/2, 2014.
- [18] P. D. Tyson and R. A. Preston-Whyte, *The Weather and Climate of Southern Africa*. Oxford, U.K.: Oxford Univ. Press, 2000.
- [19] Altitude Maps, "Global Digital Elevation Model," [Online]. Available: <https://www.altitude-maps.com/>.
- [20] R. J. Nicholls and C. Small, "Improved understanding of coastal processes and management of coastal regions," *Ocean & Coastal Management*, vol. 45, no. 6–7, pp. 527–548, 2002. doi: 10.1016/S0964-5691(02)00083-3.
- [21] C. J. C. Reason and M. Rouault, "ENSO-like decadal variability and South African rainfall," *Geophysical Research Letters*, vol. 29, no. 13, pp. 34-1–34-4, 2002. doi: 10.1029/2001GL014769.
- [22] C. J. C. Reason and D. Jagadheesha, "A model investigation of recent ENSO impacts over southern Africa," *Meteorology and Atmospheric Physics*, vol. 89, no. 1, pp. 181–205, 2005. doi: 10.1007/s00703-005-0128-9.
- [23] S. E. Falodun and B. Y. Lawal, "Investigation of non-standard refraction in a coastal area of Nigeria using radiosonde data," *FUTA Journal of Research in Sciences*, vol. 11, no. 2, pp. 358–368, 2015.
- [24] M. O. Asiyo, "Characterization and modelling of effects of clear air on multipath fading in terrestrial links," M.S. thesis, Dept. of Electrical and Electronic Eng., Univ. of Kwazulu-Natal, South Africa, 2013.