Engineering and Technology Journal e-ISSN: 2456-3358

Volume 10 Issue 01 January-2025, Page No.- 3634-3642

DOI: 10.47191/etj/v10i01.21, I.F. – 8.227

© 2025, ETJ

Neighbourhood Component Regression Approach for Housing Unit Price Prediction

Paul Boye[1], Yao Yevenyo Ziggah[2]

[1]Faculty of Engineering, Department of Mathematical Sciences, University of Mines and Technology (Ghana)

[2]Faculty of Geosciences and Environmental Studies, Department of Geomatic Engineering, University of Mines and Technology (Ghana)

ABSTRACT: Predicting housing unit price (HUP) is important for potential buyers and investors to make informed decisions. This study proposes a novel HUP prediction model based on neighbourhood component regression (NCR). The proposed NCR model was compared with other competitive methods such as principal component regression (PCR), multiple linear regression (MLR), partial least squares regression (PLSR), and generalised linear model (GLM). When tested with real datasets, the proposed NCR method revealed prediction superiority over the four state-of-the-art methods (PCR, MLR, PLSR, and GLM). This was evident from the Mean Absolute Percentage Error (MAPE), Correlation Coefficient (R), Scatter Index (SI), and Percentage Root Mean Square Error (PRMSE) utilised as model evaluation metrics. The results revealed that the NCR model had the lowest MAPE (0.0977), SI (0.0011), PRMSE (0.1130), and highest R (0.9999) as compared with the other investigated methods. This confirms the proposed NCR method's strength for efficient and reliable HUP prediction.

KEYWORDS: Neighbourhood Component Regression, Real Estate Market, Housing Unit Price Prediction

1. INTRODUCTION

Housing which is considered a portion of a sustainable economy is an important subbranch of the real estate market. In many countries, possessing a real estate property is perceived as having a high social status; and this has become the working-class goal. On the contrary, investors who are attracted by the housing market perceive real estate as an asset where money can be allocated to bring capital income; but not as a consumption good only (Hacıevliyagil et al., 2022).

For decades, researchers have extensively carried out studies in developing predictive models for housing unit price (HUP), of which data, variables, and methods implemented are usually different. Some of the well-known methods in the literature include but are not limited to multiple linear regression (MLR) (Zhang, 2021), time series analysis (Boye et al., 2019), and cost methodology (Cunha and Lobão, 2021). These methods have the benefit of being easily understood, long history application, deep-rooted acceptance, and simplicity of application due to the availability of software. It is important to mention that principal component regression (PCR) has also been used by several researchers (Gupta and Kabundi (2010); Yingying and Dongxiao, 2010) to advance HUP prediction in different jurisdictions. Literature indicates that the PCR is a feasible and effective method that could predict the HUP satisfactorily. In Tao (2019), the partial least squares regression (PLSR) method was also employed to predict HUP. A similar work was also conducted by Bork and Møller (2018) and Phan (2018). These related studies have shown that the PLSR is capable of adequately predicting the HUP. The efficiency of generalised linear model (GLM) has also been explored to predict HUP with high prediction performance (Li, 2022). Although these mentioned regression methods (MLR, PLSR, PCR, and GLM) dominate the literature in HUP prediction with satisfactory results their implementation depends on firm assumptions that are hardly met in real-world situations leading to limited generalisation performance (Shang et al., 2019).

For instance, the PCR may end up in information loss when mapped to lower dimensions (Bulut and Alma, 2011; Karamizadeh et al., 2013; Jolliffe and Cadima, 2016). The MLR models, on the other hand, perform well only when assumptions like multivariate normality, multicollinearity, and linear relationship among predictor variables are satisfied (Stigler, 1986; Gong et al., 2018). The PLSR finds it challenging to accurately eliminate redundancy and noise from the dataset to obtain more useful information to enhance its prediction robustness (Goodhue et al., 2012; Wentzell and Montoto, 2003). As useful as the GLM is, it suffers the drawback of imposition of a static model, which implies fixed relationships across observations. This restrictive assumption which contradicts reality often leads to inflexibility in modeling (West et al., 1985). The limitations posed by the mentioned methods can be overcome using neighbourhood component regression (NCR).

The NCR as a supervised dimensionality reduction technique, can facilitate the HUP prediction process by ensuring proper input variable selection and improving generalisation performance and execution speed. The NCR leverages its adaptive strength and feature selection capability to adequately learn the data (Shang et al. (2019). These characteristic features prevent the NCR method from getting overloaded and simplify the computational process. In addition, the NCR method is not subject to any specific assumptions and data conditions; and it has no information loss during the process of dimension reduction.

Therefore, the main objective of this research is to explore the capability of the NCR approach for the first time in HUP prediction. The reason is that among the plethora of advantages offered by NCR, the approach is yet to be explored and tested in the domain knowledge of HUP even though it has been applied and evaluated in diverse disciplines (e.g., medicine, hydrology, and aviation) with promising results (Tuncer and Ertam, 2020; Durocher et al., 2016; De Rivas et al., 2017). Hence, the authors deemed it fit to expand the frontiers of NCR application by implementing it in the housing industry.

The efficiency of the developed NCR model was assessed by comparing it with state-of-the-art methods of PCR, MLR, PLSR, and GLM. So, this study's main contributions to previous works are to:

- Examine and explore the capability of NCR as a new HUP prediction model; and
- Evaluate the prediction performance of NCR against PCR, MLR, PLSR, and GLM.

This research has therefore proposed a new regression methodology for improving HUP prediction accuracy.

2. METHODS

2.1 Principal Component Regression

The HUP model was developed by employing principal components analysis. Thus, the model used a multivariate technique to transform correlated variables into a few uncorrelated linear variables called principal components (PCs) (Choi et al., 2019). The first PC contains the highest amount of information. The PCR algorithm transforms the correlated variables into a smaller number of linearly combined variables of the original variables. This applied regression methodology allows for association discovery between variables and reducing their number to ease their analysis and interpretation. Equation (1) shows the values of the predictor variables PCs for each observation (Figueroa-Garcia et al., 2021; Jolliffe, 2011).

$$\mathbf{U} = \mathbf{ZV} \tag{1}$$

where **U** is a matrix whose $(i,k)^{th}$ element represents the k^{th} PC value in observation i^{th}. Matrix **Z** with $(n \times p)$ dimension, its $(i,j)^{th}$ element represents the value of the j^{th} predictor variable in i^{th} observation. **V** is a matrix whose k^{th} column is the unit eigenvector associated with k^{th} greater eigenvalue of $\frac{1}{n}\mathbf{Z}'\mathbf{Z}$; where $\mathbf{Z}'$ is the centered and scaled matrix **Z**.

The PCR method is an estimation technique, its mathematical equation is given as Equation (2).

$$g = \beta_0 + \sum_{i=1}^{n} \beta_i U_i \tag{2}$$

where g is the dependent variable, β_0 is the intercept, β_i is the component coefficient, and U_i represents the PC.

2.2 Multiple Linear Regression

Considering multiple variables and one outcome dataset $\left(t_{i1}\ t_{i2}\ ....\ t_{i,p-1}; g_i\right)$ for $i = 1, 2, ..., n$ units of observations, MLR which formalises a simultaneous statistical relation between a single continuous outcome g and the predictor variables $T_r\ (r = 1, 2, ..., p-1)$ (Equation (3)) is an extension of the simple linear regression (Labban, 2020).

$$g_i = \beta_0 + \sum_{r=1}^{j} \beta_r t_{ir} + \varepsilon_i \tag{3}$$

where β_0 is the intercept on the g_i axis. That is the average of *g* when all $t_r = 0.$. Each β_r represents the average with respect to t_r. That is, the magnitude of the change in the average of *g* when t_r is larger by one unit when all the other predictors are held constant. ε_i is the random error of the i^{th} observation for $i = 1, ..., n.$ Equation (3) can be written more compactly as shown in Equation (4):

$$\mathbf{G} = \mathbf{T\beta} + \mathbf{U} \tag{4}$$

where **T** is the vector having dimension of the *n* dependent observations, **T** is the independent (explanatory) matrix of dimension $n \times (r+1)$. $\boldsymbol{\beta}$ is the regression coefficients vector of dimension $(j+1)$, and $\mathbf{U} \sim N\left(0, \sigma^2\right)$ is the vector of random errors of dimension n. σ^2 is the population variance.

Using the maximum likelihood estimation method, the likelihood function (Equation (6)) of the function in Equation (5) which is from the normal distribution is given as (Jäntschi et al., 2016):

$$f\left(\mathbf{U}_i\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{U}_i^2\right) \tag{5}$$

$$L\left(\mathbf{U}, \boldsymbol{\beta}, \sigma^2\right) = \prod_{i=1}^{n} f\left(\mathbf{U}_i\right) = \sigma^{-n}\left(2\pi\right)^{-n/2} \exp\left(-\frac{1}{2}\sigma^{-2}\mathbf{U}'\mathbf{U}\right) \tag{6}$$

Since $\mathbf{U} = \mathbf{G} - \mathbf{T\beta}$, the likelihood function becomes

$$L\left(\mathbf{G},\mathbf{T},\boldsymbol{\beta},\sigma^2\right)=\sigma^{-n}\left(2\pi\right)^{-n/2}\exp\left(-\frac{1}{2}\sigma^{-2}\left(\mathbf{G}-\mathbf{T}\boldsymbol{\beta}\right)'\left(\mathbf{G}-\mathbf{T}\boldsymbol{\beta}\right)\right) \tag{7}$$

Taking the natural log of the likelihood function gives Equation (8).

$$\ln\left[L\left(\mathbf{G},\mathbf{T},\boldsymbol{\beta},\sigma^2\right)\right]=-\frac{n}{2}\ln\left(2\pi\right)-n\ln\left(\sigma\right)-\frac{1}{2}\sigma^{-2}\left(\mathbf{G}-\mathbf{T}\boldsymbol{\beta}\right)'\left(\mathbf{G}-\mathbf{T}\boldsymbol{\beta}\right) \tag{8}$$

By differentiating Equation (9) with respect $\boldsymbol{\beta}$ and equate to zero, results:

$$\left(\mathbf{G}-\mathbf{T}\boldsymbol{\beta}\right)'\left(\mathbf{G}-\mathbf{T}\boldsymbol{\beta}\right)=0 \tag{9}$$

$$\mathbf{G}'\mathbf{G}-2\boldsymbol{\beta}'\mathbf{T}'\mathbf{G}+\boldsymbol{\beta}'\mathbf{T}'\mathbf{T}\boldsymbol{\beta}=0 \tag{10}$$

To find the value of $\boldsymbol{\beta}$ that minimises the model errors as much as possible, derivate with respect to $\boldsymbol{\beta}$ is taken:

$$-2\mathbf{T}'\mathbf{G}+2\mathbf{T}'\mathbf{T}\hat{\boldsymbol{\beta}}=0 \tag{11}$$

$$\mathbf{T}'\mathbf{G}=\mathbf{T}'\mathbf{T}\hat{\boldsymbol{\beta}} \tag{12}$$

$$\hat{\boldsymbol{\beta}}=\left(\mathbf{T}'\mathbf{T}\right)^{-1}\mathbf{T}'\mathbf{G} \tag{13}$$

$\hat{\boldsymbol{\beta}}$ is the vector of predictors.

2.3 Partial Least Squares Regression

For studies with strongly correlated predictor variables, PLSR is a technique often applied in the case of multivariate regression (Wold, 1975; Wold et al., 2001). The technique is also considered as latent variable regression. Thus, the technique extracts latent explanatory variables from the original set of correlated variables. Iteratively, the PLSR algorithm constructs an orthonormal latent components sequence (basic vectors) from the explanatory variables having maximal covariance with the response variable. Consequently, based on the latent variables, the regression vector is computed, and the process eliminates multicollinearity difficulties. In each iteration, a linear model (Equation (14)) is fitted, and the solution vectors and the dataset are projected onto low dimension subspaces.

$$\mathbf{y}=\mathbf{X}\beta+\varepsilon \tag{14}$$

where β is regression vector to be estimated and ε is a constant error. The error is identically distributed having constant variance and zero expectation.

2.4 Generalised Linear Model

GLM which was proposed first by Nelder and Wedderburn (1972) provides a framework for relating response and predictor variables. The GLM is a flexible extension of linear regression. The method permits a nonlinear relationship between the explanatory and the response variables of a linear model, and diverse data generation processes (Khuri et al., 2006; Paul et al., 2013). For a given random variable y, Equation (15) shows the probability density function (James, 2002; Dunn and Smyth, 2018).

$$f\left(y,\theta,\phi\right)=exp\left[\frac{y\theta-b\left(\theta\right)}{a\left(\phi\right)}+c\left(y,\phi\right)\right] \tag{15}$$

where θ is the canonical parameter; and it is the parameter of interest. $a(.)$, $b(.)$, and $c(.)$ are given functions. ϕ (constant) is the scale parameter. The linear regression relationship between the predictor **X** and the response *Y* is given in Equation (16) as

$$g\left(\mu\right)=\beta_0+\sum_{j=1}^{p}\beta_j x_j \tag{16}$$

$$\text{where } \mu=E\left(Y;\theta,\phi\right)=b'\left(\theta\right) \tag{17}$$

g is known as the link function, β_0 represents the intercept. Each β_j denotes the slope associated with X_j. GLM provides a very flexible class of procedures. However, they assume that the predictor has a finite dimension.

2.5 Neighbourhood Component Regression

The NCA is a nonparametric technique for choosing relevant features to maximise regression model prediction accuracy, using a constructed complete graph with each data point serving as its node. Consider a dataset *G* comprising *H* data points which is given as $G=\{u_i,v_i\}$ for $i=1,2,...,H$. In learning distance metric, some form of supervision information is aimed at learning Mahalanobis matrix *B*. Mahalanobis distance metric (Equation (18)) is the distance squared between two data points u_i. and v_i (Wang and Tan, 2017).

$$g^2{}_B\left(u_i,v_{i,}\right)=\left(u_i-v_j\right)'B\left(u_i-v_j\right) \tag{18}$$

where $B\geq 0$ is a semidefinite positive matrix. $u_i,v_j\in R^g$ is a pair of samples. Between any two nodes, let the weight of each edge be represented as T_{ij}; which is interpreted as the probability that the data point u_i selects u_j as its neighbour and can be computed as shown in Equation (19) as:

$$T_{ij}=\frac{\exp\left(-g^2{}_{B^{ij}}\right)}{\sum_{t\in H_i}\exp\left(-g^2{}_{B^{it}}\right)} \tag{19}$$

where H_i denotes the set of neighbours of u_i. If $T_{ij}\geq 0$ and $\sum_{j\in H_i}T_{ij}=1$, then T_{ij} is a valid probability measure.

The NCA objective is to learn a linear transformation B (Equation (20)) that maximises the log-likelihood, that after transformation each data point selects the points with the same labels as itself as neighbours.

$$\max L(B) = \sum_i \log\left(\sum_{j \in H_i} 1(v_i = v_j).T_{ij} \right) \tag{20}$$

3. MODEL PREDICTION EVALUATION PERFORMANCE

The developed regression model efficiency for HUP prediction was determined in this study by the following statistical indicators: Mean Absolute Percentage Error (MAPE), Correlation Coefficient (CC), Scatter Index (SI), and Percentage Root Mean Square Error (PRMSE) (Chicco et al., 2021; Li and Liu, 2020; Lin et al., 2016).

3.1 Mean Absolute Percentage Error

This index has a very intuitive interpretation in terms of relative error. The MAPE (Equation (21)) indicates an average of the absolute percentage of errors. The lower the MAPE value relative to the actual data, the higher the accuracy of the developed model.

$$MAPE = m^{-1} \sum_{i=1}^{m} \left| \frac{O_i - P_i}{O_i} \right| \times 100 \tag{21}$$

where in Equations (21) - (22) and (24) *m* is the total number of test samples, O_i and P_i represent the observed and predicted values respectively.

3.2 Correlation Coefficient

This index elucidates the inconsistency in the predicted values and their relation to the observed values. The ideal values lie between 0 and 1. Values closer to 1 indicate a better fit. The CC is expressed in Equation (22).

$$CC = m^{-1} \left[\frac{\sum_{i=1}^{m} (P_i - \bar{P})(O_i - \bar{O})}{\sigma_p \sigma_o} \right] \tag{22}$$

where $\bar{O}$ and $\bar{P}$ represent the mean of the observed and predicted values respectively. σ_o and σ_p are the standard deviations of the observed and the predicted values respectively.

3.3 Percentage Root-Mean Square Error

The PRMSE (Equation (23)) is a scale-independent measure that evaluates the accuracy of a model's predictive performance.

$$\text{PRMSE} = \frac{RMSE}{\bar{x}} \times 100 \tag{23}$$

where $\bar{x}$ is the mean value of the observations, and

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} (O_i - P_i)^2}{m}} \tag{24}$$

where *SD* represented the actual observation and the prediction values standard deviation, coverage factor of 1.96 agreed to a 95% confidence level. Root-Mean Square Error is represented by RMSE.

3.4 Scatter Index

SI is a normalized measure of error often reported as a percentage. Lower values of SI indicate better model performance, and they are computed mathematically, as shown in Equation (25).

$$\text{SI} = \frac{RMSE}{\bar{x}} \tag{25}$$

Where $\bar{x}$ is the mean value of the observations?

4. NUMERICAL APPLICATION

4.1 Data Used

The fifteen-year dataset for a one-bedroom housing unit from Regimanuel Gray Estates Ltd., an estate development corporation in Ghana, West Africa, was used for the model development. The dataset (cement, sand, iron rods, roofing, paint, and wood) served as independent variables, and half-yearly HUP (dependent variable), spans from 2003 to 2017. In total, 30 observations were applied. Table 1 presents the statistical summary of the dataset used for the regression work.

Table 1: Statistical Summary of One-Bedroom HUP Dataset

Parameter	Mean Value ($)	Median Value ($)	Minimum Value ($)	Maximum Value ($)	Standard Deviation ($)
Cement	2486.25	2109.89	60.59	6119.39	2034.56
Sand	8949.44	10328.17	1467.16	15339.66	4180.64
Iron Rods	560.53	560.83	319.52	819.10	151.08
Roofing	3395.84	3418.23	1266.55	5464.05	1301.38
Paint	801.46	802.13	5.75	1592.75	1592.75
Wood	970.71	970.79	280.33	1661.63	424.99
HUP	55225.48	52602.50	31455.00	83600.00	15011.80

4.2 Models Formulated

To develop the various regression models, 24 data points (training data) were utilised. For model validation, 6 data points (testing data) were employed. In this study, five different regression models, namely MLR, PLSR, GLM, NCR, and PCR were developed. The formulated MLR model is given in Equation (26).

$$\begin{aligned}\mathrm{HUP}_{\mathrm{MLR}} &= 34452.18801 + (0.5210\times \mathrm{cement}) - (0.5184\times \mathrm{sand}) + (24.2465\times \mathrm{ironrods}) - \\ &(14.5267\times \mathrm{roofing}) + (37.5979\times \mathrm{paint}) + (30.5891\times \mathrm{wood})\end{aligned} \tag{26}$$

The PLSR model developed for predicting HUP is expressed in Equation (27).

$$\begin{aligned}\mathrm{HUP}_{\mathrm{PLSR}} &= 22135.86166 + (1.912741291\times \mathrm{cement}) - (0.300148713\times \mathrm{sand}) + \\ &+ (0.915448285\times \mathrm{ironrods}) + (7.594306744\times \mathrm{roofing}) + (2.849479293\times \mathrm{paint}) \\ &+ (2.465646748\times \mathrm{wood})\end{aligned} \tag{27}$$

From Figure 1, the optimal number of components is 3. Hence, the percentage variance explained in the HUP prediction by the model with the 3 components is greater than 99%; and this confirms the robustness of the choice made.

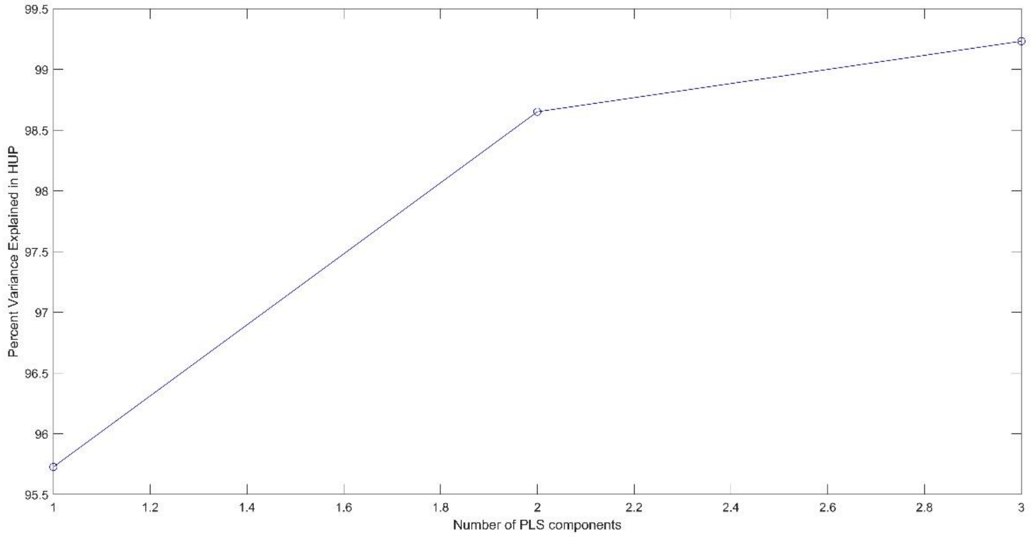

Figure 1: Percent Variance Explained as a Function of Number of PLS Components

Employed in the Analysis

The final GLM model is expressed in Equation (28).

$$\begin{aligned}\mathrm{HUP}_{\mathrm{GLM}} &= 34452.18801 + (0.5210\times \mathrm{cement}) - (0.5184\times \mathrm{sand}) + (24.2465\times \mathrm{ironrods}) \\ &- (14.5267\times \mathrm{roofing}) + (37.5979\times \mathrm{paint}) + (30.5891\times \mathrm{wood})\end{aligned} \tag{28}$$

In the application of the NCR methodology, a distinction between redundant and relevant input features is carried out during the model building. A feature is deemed relevant when its weight exceeds one. Features with zero weights are considered to have little impact on the response features and thus excluded. Figure 2 shows the three selected dependent features, cement, sand, and roofing, whose feature weights exceed one and thus was used as the input variables in the NCR model development. Figure 3 shows the schematic diagram for the NCA regularisation parameter (λ) results.

The optimal λ value for the selected three features is 28496.98 with an MSE value of 549.14. The optimisation algorithm used to fine tune the regularization parameter is the mini-batch-Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm. The gradient tolerance was set at 1×10^{-4} with an iteration limit of 1000.

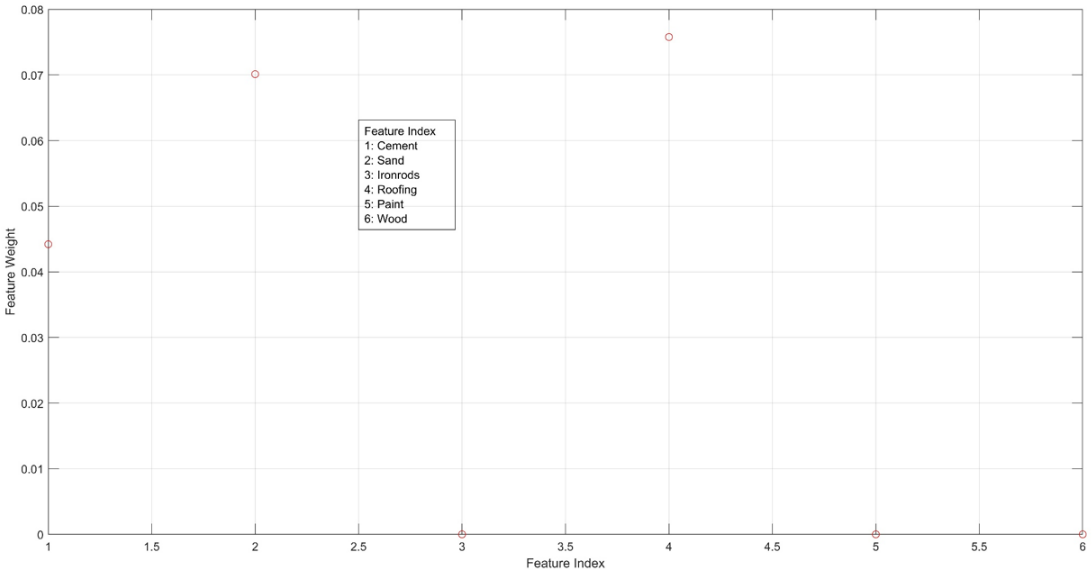

Figure 2: Feature Selection Results for NCA

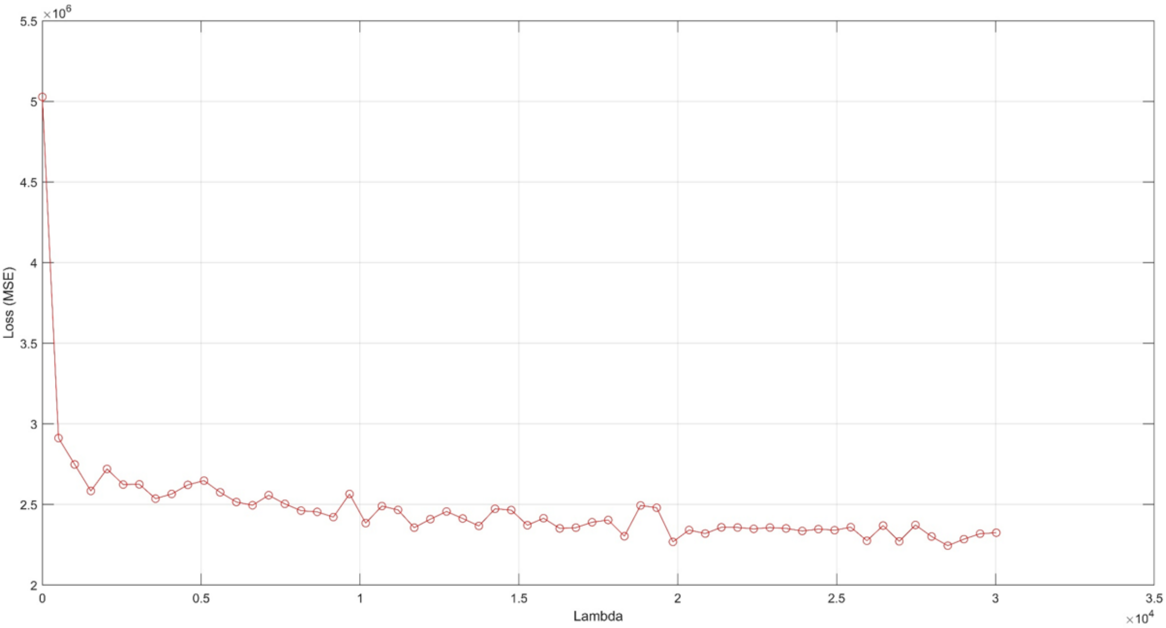

Figure 3: Result for Mean Loss Function and Regularisation Parameter

The PCR model formed is expressed in Equation (29).

$$\begin{aligned}\mathrm{HUP}_{\mathrm{PCR}} &= 29522.9284 + (4.130063053 \times \mathrm{cement}) - (1.120358574 \times \mathrm{sand}) + \\ &(0.191689931 \times \mathrm{ironrods}) - (1.354440398 \times \mathrm{roofing}) + (0.530580747 \times \mathrm{paint}) \\ &+ (0.463840007 \times \mathrm{wood})\end{aligned} \tag{29}$$

4.3 Developed Model Efficiency Test Results

During the development of the regression models (PCR, MLR, PLSR, GLM, and NCR), the acquired dataset was divided into an 80% training set (24 data points) and a 20% testing set (6 data points). The model fit was done using the training set, while the testing set served as independent data to validate the forecasting strength of the developed models. Table 2 shows the intercomparison among the applied methods.

Table 2: Statistical Analysis Testing Results

Model	MAPE	R	PRMSE	SI
PCR	1.8646	0.9980	2.0873	0.0209
MLR	1.3390	0.9987	1.6411	0.0164
PLSR	1.2441	0.9987	1.5029	0.0150

GLM	1.3390	0.9987	1.6411	0.0164
NCR	0.0977	0.9999	0.1131	0.0011

From Table 2, the R statistic results indicate the degree of linear association between the actual and the predicted HUP values for each competing method. Thus, the R-value depicts the model prediction accuracy, and the larger the value the better the agreement between the actual and the predicted values. From Table 2, the NCR method revealed a strong positive R-value of 0.9999, with GLM, PLSR, and MLR closely following with 0.9987. The PCR obtained an R of 0.9980. The interpretation is that the NCR method achieved the highest degree of closeness between its predictions and the actual HUP data.

Besides the R-value, the MAPE, PRMSE and SI indices portray the models' bias in estimating the HUP. Consequently, the lower the degree of dispersion of the values of these indices from zero, the better the model's acceptable accuracy. Thus, the proposed NCR method had the lowest MAPE value of 0.0977 followed by PLSR and GLM competing together, and PCR having the worse value of 1.8646, respectively. Based on the PRMSE results, it is obvious that the NCR predicted HUP values are practically more acceptable than the other investigated methods. A similar assertion was observed for the SI values where the NCR produced 0.0011 as compared to the GLM, PLSR, MLR and PCR. Based on the statistical results it can be inferred that the NCR is the most superior in predicting HUP. Hence, the NCR can be regarded as an effective method for predicting HUP.

5. CONCLUSIONS

This study has developed five regression models (PCR, MLR, PLSR, GLM, and NCR) to predict HUP based on case study data obtained from Ghana, West Africa. For the first time, the study proposes a new HUP prediction model-based NCR. The proposed NCR method relied on its feature selection capability to improve HUP prediction accuracy. Generally, results of the statistical analysis revealed that the developed regression models are good and can be used to predict HUP based on their performance indicators. However, the proposed NCR method is the most suitable for predicting HUP based on its achieved low MAPE (0.0977%), PRMSE (0.1130%), SI (0.0011), and high R (0.9999) values as compared to the other contending methods. The proposed NCR method will be advantageous not only to potential buyers and investors but to stakeholders in policymaking and the housing industry as well as in developing countries like Ghana where planning and making policies about the availability of housing is paramount.

DATA AVAILABILITY

The housing unit price data used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

1. Bork, L. and Møller, S.V. (2018), "Housing price forecastability: A factor analysis", *Real Estate Economics*, Vol. 46 No. 3, pp.582-611.
2. Boye, P., Mireku-Gyimah, D. and Sadiq, H. (2019), "Time series analysis model for estimating housing unit price", *Ghana Journal of Technology*, Vol. 3 No., pp.35-41.
3. Bulut, E. and Alma, Ö.G. (2011), "Dimensionality reduction methods: PCR, PLSR, RRR and a health application", *Physical Sciences*, Vol. 6 No. 2, pp.36-47.
4. Chicco, D., Warrens, M.J. and Jurman, G. (2021), "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation", *PeerJ Computer Science*, Vol. 7, p.e623.
5. Choi, Y.Y., Shon, H., Byon, Y.J., Kim, D.K. and Kang, S. (2019), "Enhanced application of principal component analysis in machine learning for imputation of missing traffic data", *Applied Sciences*, Vol. 9 No. 10, p.2149.
6. Cunha, A.M. and Lobão, J. (2021), "The determinants of real estate prices in a European context: a four-level analysis", *Journal of European Real Estate Research*.
7. De Rivas, B.L., Vivancos, J.L., Ordieres-Meré, J. and Capuz-Rizo, S.F. (2017), "Determination of the total acid number (TAN) of used mineral oils in aviation engines by FTIR using regression models", *Chemometrics and Intelligent Laboratory Systems*, Vol. 160, pp.32-39.
8. Despotovic, M., Nedic, V., Despotovic, D. and Cvetanovic, S. (2016), "Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation", *Renewable and Sustainable Energy Reviews*, Vol. 56, pp.246-260.
9. Dunn, P.K. and Smyth, G.K. (2018), Generalized Linear Models with Examples in R, Vol. 53, New York: Springer.
10. Durocher, M., Chebana, F. and Ouarda, T.B. (2016), "Delineation of homogenous regions using hydrological variables predicted by projection

pursuit regression", *Hydrology and Earth System Sciences*, Vol. 20 No. 12, pp.4717-4729.

11. Figueroa-Garcia, E., Segura-Castruita, M.A., Luna-Olea, F.M., Vázquez-Vuelvas, O.F. and Chávez-Rodríguez, A.M. (2021), "Design of a hybrid solar collector with a flat plate solar collector and induction heating: evaluation and modeling with principal components regression", *Revista Mexicana de Ingeniería Química*, Vol. 20 No. 3, pp. Alim2452-Alim2452.
12. Goldberger, J., Hinton, G.E., Roweis, S.T. and Salakhutdinov, R. (2004), "Neighbourhood components analysis", Advances in Neural Information Processing Systems, pp. 513–520.
13. Gong, Z., Liu, C., Sun, J., and Teo, K.L. (2018), "Distributionally robust L1-estimation in multiple linear regression", *Optimization Letters*, pg.1-13.
14. Goodhue, D.L., Lewis, W. and Thompson, R. (2012), "Does PLS have advantages for small sample size or non-normal data?", *MIS Quarterly*, pp.981-1001.
15. Gupta, R. and Kabundi, A. (2010), "Forecasting real US house prices: Principal components versus Bayesian regressions", *International Business & Economics Research Journal (IBER)*, Vol 9 No. 7.
16. Hacıevliyagil, N., Drachal, K. and Eksi, I.H. (2022), "Predicting house prices using DMA method: Evidence from Turkey", *Economies*, Vol. 10 No. 3, p.64.
17. James, G.M. (2002), "Generalized linear models with functional predictors", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 64 No. 3, pp.411-432.
18. Jäntschi, L., Bálint, D. and Bolboacă, S. (2016), Multiple linear regressions by maximizing the likelihood under the assumption of generalized Gauss-Laplace distribution of the error", *Computational and Mathematical Methods in Medicine*, Vol. 2016, pp.1-8.
19. Jolliffe, I. (2011), *Principal Component Analysis*, Springer: New York, NY, USA.
20. Jolliffe, I.T. and Cadima, J. (2016), "Principal component analysis: a review and recent developments", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 374 No. 2065, p.20150202.
21. Karamizadeh, S., Abdullah, S.M., Manaf, A.A., Zamani, M. and Hooman, A. (2013), "An overview of principal component analysis", *Journal of Signal and Information Processing*, Vol. 4 No. 3B, p.173.
22. Khuri, A.I., Mukherjee, B., Sinha, B.K. and Ghosh, M. (2006), "Design issues for generalized linear models: A review", *Statistical Science*, Vol. 21 No. 3, pp.376-399.
23. Labban, J.A. (2020), "Estimating multiple linear regression parameters using term omission method", *Periodicals of Engineering and Natural Sciences (PEN)*, Vol. 8 No. 4, pp.2290-2299.
24. Li, M. and Liu, K. (2020), "Probabilistic prediction of significant wave height using dynamic Bayesian network and information flow", *Water*, Vol. 12 No. 8, p.2075.
25. Li, M.F., Tang, X.P., Wu, W. and Liu, H.B. (2013), "General models for estimating daily global solar radiation for different solar radiation zones in mainland China", *Energy Conversion and Management*, Vol. 70, pp.139-148.
26. Li, X. (2022), Prediction and analysis of housing price based on the generalised linear regression model", *Computational Intelligence and Neuroscience*.
27. Lin, C., Thomson, G. and Popescu, S.C. (2016), "An IPCC-compliant technique for forest carbon stock assessment using airborne LiDAR-derived tree metrics and competition index", *Remote Sensing*, Vol. 8 No. 6, p.528.
28. Nelder, J.A. and Wedderburn, R.W. (1972), "Generalised linear models", *Journal of the Royal Statistical Society: Series A (General)*, Vol. 135 No. 3, pp.370-384.
29. Paul, G., Cardinale, J. and Sbalzarini, I.F. (2013), "Coupling image restoration and segmentation: a generalized linear model/Bregman perspective", *International journal of computer vision*, Vol. 104 No. 1, pp.69-93.
30. Phan, T.D. (2018), "Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia", *IEEE in 2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, pp.35-42.
31. Shang, Q., Tan, D., Gao, S. and Feng, L. (2019), "A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis", *Journal of Advanced Transportation*.
32. Stigler, S.M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, The Belknap Press of Harvard University Press, Cambridge.
33. Tao, Q. (2019), "Analysis of commodity housing price based on partial least squares regression", *Academic Journal of Computing & Information Science*, Vol. 2 No. 3.
34. Tuncer, T. and Ertam, F. (2020), "Neighborhood component analysis and relief based survival recognition methods for Hepatocellular carcinoma", *Physica A: Statistical Mechanics and its Applications*, Vol. 540, p.123143.

35. Wang, D. and Tan, X. (2017), “Bayesian neighborhood component analysis”, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29 No. 7, pp.3140-3151.
36. Wentzell, P.D. and Montoto, L.V. (2003), “Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures”, *Chemometrics and Intelligent Laboratory Systems*, Vol. 65 No. 2, pp.257-279.
37. West, M., Harrison, P.J., and Migon, H.S. (1985), “Dynamic generalized linear models and Bayesian forecasting”, *Journal of the American Statistical Association*, Vol. 80 No. 389, pp.73-83.
38. Wold, S., Sj ̈ostr ̈om, M. and Eriksson, L. (2001), “PLS-regression: a basic tool of chemometrics”, *Chemometrics and Intelligent Laboratory Systems*, Vol. 58 No. 2, pp.109–130.
39. Wold, H. (1975), “Soft modeling by latent variables: The non-linear iterative partial least squares (NIPALS) approach”, *Journal of Applied Probability*, Vol. 12 No. S1, pp.117-142.
40. Yingying, L. and Dongxiao, N. (2010), “Application of principal component regression analysis in power load forecasting for medium and long term”, *IEEE in 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, Vol. 3, pp. V3-201.
41. Zainuri, N.A., Jemain, A.A. and Muda, N. (2015), “A comparison of various imputation methods for missing values in air quality data”, *Sains Malaysiana*, Vol. 44 No. 3, pp.449-456.
42. Zhang, Q. (2021), “Housing price prediction based on multiple linear regression”, *Scientific Programming*.