Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 16(2) 2023: 170-181
http://ejournal.undip.ac.id/index.php/media_statistika

PER CAPITA CONSUMPTION ESTIMATION IN SURABAYA USING ENSEMBLE MODEL APPROACH

Sutikno, Jerry Dwi Trijoyo Purnomo, Unggul Harfianto, Yoga Prastya Irfandi, Kartika Nur Anisa, Fajar Dwi Cahyoko
Department of Statistics, ITS Surabaya, Indonesia

e-mail: *sutikno@statistika.its.ac.id*

DOI: 10.14710/medstat.16.2.170-181

Article Info:
Received: 14 September 2023
Accepted: 19 February 2024
Available Online: 28 February 2024

Keywords:
Proxy Mean Test; Per Capita Consumption; Ensemble Model

Abstract: The categorization of the Low-Income Community category is based on the poverty indicators in the Multidimensional Poverty Index, including the dimensions of health, education, and living standards. The Proxy Means Test (PMT) can estimate household income or consumption by taking into account household conditions that are readily observable and cannot be manipulated. This method offers the advantage of being capable of determining both the poverty level of a household and the household's characteristics based on asset ownership and socio-demographic conditions. This study aims to estimate per capita consumption using OLS, Robust, Quantile, LASSO, and Ensemble methods. The application of these methods is intended to address various issues, including the presence of outlier data, multicollinearity, and uncertainties. The results indicate that none of the four methods used achieved the highest accuracy based on the MSE, MAE, and sMAPE criteria. Consequently, employing an ensemble model becomes essential to accommodate the element of uncertainty present in these four models. The application of the ensemble method is not only as a comparison between the models, but also as a means to capture the uncertainty contained in each model

1. INTRODUCTION

Surabaya City is currently making concerted efforts to alleviate poverty and enhance the welfare of its residents. In 2020, the poverty rate in Surabaya City stood at 5.02 percent, marking a 0.51 percent increase from the previous year (BPS, 2021). One of the initiatives undertaken by the Surabaya City government is the establishment of a poverty alleviation program tailored to the specific characteristics of impoverished households.

For the determination of poor household targets, the Surabaya City Government has used the Proxy Means Test (PMT) method (Surabaya Mayor Regulation Number 58 of 2019). The PMT method has been adopted by the National Team for the Acceleration of Poverty Reduction (TNP2K) in selecting poor households receiving social assistance or programs since 2005. The PMT method can estimate household income or consumption by considering household conditions that are observable and cannot be manipulated. This method has the advantage of being able to determine the poverty level of a household, as well as the characteristics of the household according to asset ownership and socio-demographic conditions. Furthermore, appropriate programs for poverty alleviation can be

determined. The PMT method was developed based on regression analysis. Several studies were conducted regarding PMT in Surabaya City such as Fitri, 2019; Prabeswari, 2019; Malta, 2019; and Harfianto, 2021. The regression methods used are multiple regression, quantile regression, robust ridge regression, tree regression, and Least Absolute Shrinkage and Selection Operator (LASSO). In the process of preparing the regression model, there are several problems including the presence of outlier data and multicollinearity, so the model prepared can accommodate these problems.

The accuracy of the models that have been used is quite diverse, including multiple regression, quantile regression, and robust ridge. Each of the methods used still has an error, in other words, there is still an element of uncertainty. The modeling process depends on uncertainty. The process of modeling household expenditure predictions is complex, accommodating many components that need to be considered. The problem that is often faced is how the uncertainty can be quantified into a model by considering the sources of uncertainty. This can be solved with a combination of prediction models commonly referred to as ensemble models. The basic idea of this model combination is that each model has a different ability to capture data patterns (Zhang, 2003). Several studies examining combination models, including those by Blanc & Setzer (2016) using two models forecast, Peng et al. (2017) using Outlier Robust Extreme Learning Machine and Time-varying Mixture Copula Function , and Wang et al. (2016) using Ensemble Empirical Mode Decomposition and GA-BP neural network, have indicated that the combination method enhances the accuracy of prediction results. This research focuses on the application of the ensemble method to reduce variance and error in predicting the upcoming per capita consumption as well as capturing the element of uncertainty in the prediction model.

2. LITERATURE REVIEW

2.1. Ordinary Least Square Regression

Regression analysis is one of the techniques in statistics that is used to determine the relationship between several variables and predict a response variable (Kutner et al., 2004). Regression analysis can be interpreted as a statistical method that explains the relationship pattern (model) between two or more variables.

The model representing the relationship between the response variable (y) and the predictor variables (X) is shown in Equation (1) (Montgomery et al., 1992).

$$\boldsymbol{y}_{(n\times 1)} = \boldsymbol{X}_{(n\times(k+1))}\boldsymbol{\beta}_{((k+1)\times 1)} + \boldsymbol{\varepsilon}_{(n\times 1)} \quad (1)$$

Thus, the estimated value of $\widehat{\boldsymbol{\beta}}$ is $\widehat{\boldsymbol{\beta}} = \boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, where $\boldsymbol{\varepsilon}$ is an error vector of size (n×1) with normal distribution, no autocorrelation and homoscedasticity assumption, $\boldsymbol{y}$ is a vector of response variables of size (n×1), $\boldsymbol{X}$ is a matrix of predictor variables of size (n×(k+1)) and $\boldsymbol{\beta}$ is a vector of regression parameters of size ((k+1)×1).

2.2. Robust Regression

At the point when the observations y in a linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ are normally distributed, the least squares method performs well in the sense that it produces an estimate $\boldsymbol{\beta}$ that has good statistical properties. However, when observations follow a non-normal distribution, especially observations that do not have longer or heavier tails than normal, the least squares method may not be appropriate (Pratiwi et al., 2018). There are three classes of problems that can use robust techniques (Chen, 2002), namely: 1) Problems with outliers in the y (response) variable, 2) Problems with outliers contained in variable x (leverage point), 3) Problems with outliers contained in both y (response) and x (predictor) variables.

2.3. Robust M Estimation Method

M-estimation is recognized as a straightforward approach in robust regression. The basic idea is to minimize the objective function as in Equation (2) (Pratiwi et al., 2018)

$$\sum_{i=1}^{n} \rho(e_i^*) = \sum_{i=1}^{n} \rho\left(\frac{\boldsymbol{e_i}}{\hat{\boldsymbol{\sigma}}}\right) = \sum_{i=1}^{n} \rho\left(\boldsymbol{y_i} - \frac{\boldsymbol{e_i}}{\hat{\boldsymbol{\sigma}}}\right) \quad (2)$$

The value of $\hat{\boldsymbol{\sigma}}$ (the scale of the robust estimate) is obtained through iteration

$$\hat{\sigma} = med_{i=1}^{n} \left| y_i - X_i^{(l-1)} \right| / \beta_0$$

where l $(l = 1,2, \ldots)$ is an iteration. $\beta_0 = \Phi^{-1}(0{,}75)$ and Φ^{-1} is the inverse of the standard normal cumulative function. $\rho(e_i^*)$ is the symmetric function of the residuals or the function that contributes each residual to the objective function. $\psi = \rho'$ with ψ the derivative of ρ. $\psi(.)$ is the effect function used in deriving the weights. Given a weighting function $w_i = \frac{\psi e_i^*}{e_i^*}$ into Equation (3)

$$\sum_{i=1}^{n} w_i \left(\frac{y_i - \boldsymbol{X_i} b}{\hat{\sigma}}\right) X_i = 0 \quad (3)$$

Equation (3) is denoted in matrix form as Equation (4).

$$\boldsymbol{X^T W X b = X^T W y} \quad (4)$$

Equation (4) is referred to as weighted least squares that minimize $\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$. Weighted least squares can be employed for calculating the estimation of *M*, resulting in the parameter estimation presented in Equation (5).

$$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{M}} = (\boldsymbol{X^T W X})^{-1} \boldsymbol{X^T W y} \quad (5)$$

The weights in *M* estimation depend on the residuals and coefficients. To address this issue, an iterative procedure known as Iteratively Reweighted Least Squares (IRLS) is necessary. In this research, Tukey Bisquare is employed to estimate the parameters. The general forms of the objective function, influence function, and weight function for Tukey Bisquare are delineated in Table 1 (Khan et al., 2021).

Table 1. Objective Function, Effect Function and Weighting Function for M-Tukey Bisquare Estimation

	***Tukey Bisquare* Method**
Objective Function	$\rho_B(e^*) = \begin{cases} \frac{k^2}{6}\left[1 - \langle 1 - \left(\frac{e_i^*}{r}\right)^2 \rangle^3\right], for\ \lvert e_i^* \rvert \le r \\ r^2/6 \quad , for\ \lvert e_i^* \rvert > r \end{cases}$
Effect Function	$\psi_B(e^*) = \begin{cases} e_i^* \langle 1 - \left(\frac{e_i^*}{r}\right)^2 \rangle^2 \ for\ \lvert e_i^* \rvert \le r \\ 0 \quad , for\ \lvert e_i^* \rvert > r \end{cases}$
Weighting Function	$\psi_B(e^*) = \begin{cases} \langle 1 - \left(\frac{e_i^*}{r}\right)^2 \rangle^2 \ for\ \lvert e_i^* \rvert \le r \\ 0 \quad , for\ \lvert e_i^* \rvert > r \end{cases}$

2.4. Quantile Regression

Quantile regression is one of the development methods of linear regression, which can be properly used to overcome the deviation of the homogeneity assumption in linear regression where the residual data gets bigger and becomes inhomogeneous due to the presence of outlier data. This condition can be indicated by the shape of the data which is no longer symmetrical due to the presence of outliers. Then, the simple regression method using the average in estimating the model parameters is no longer appropriate, so the quantile regression method is used (Amédée-Manesme et al., 2020). Suppose *Y* is a random variable with distribution function F_Y and quantiles $(\tau) \in (0,1)$ can be expressed as $F_Y(y) = F(y) =$

$P(Y \leq y)$ and the inverse function is denoted as $Q_Y(\tau) = F_Y{}^{-1}(\tau) = inf\{y: F_Y(y) \geq \tau\}$. If Y is a known function of X, its probability functions are expressed as $F_{Y|X}(y)$ and $Q_{Y|X}(\tau)$. The linear model of the quantile regression equation of the probability function can be denoted as Equation (6) (Hong et al., 2019).

$$Q_{Y|X}(\tau) = \alpha(\tau) + \beta(\tau)x \tag{6}$$

The quantile regression model, when the X variable is a binary category can be expressed in Equation (7)

$$Q_{Y|D}(\tau) = \alpha(\tau) + \delta(\tau)D \tag{7}$$

where D denotes the categorical variable, with D = 1 as the category under study and D = 0 as a control. The generalized conditional quantile regression equation $Q_{Y|X}(\tau)$ in Equation (6) can be explained as in Equation (8).

$$y_i = \beta_{\tau 0} + \beta_{\tau 1}x_{1k} + \cdots + \beta_{\tau k}x_{ik} + \varepsilon_{\tau i}, i = 1,2,\cdots,n \tag{8}$$

where y_i is response variable with i-th observation, x_{ik} is k-th predictor variable and i-th observation, $\beta_{\tau k}$ is parameters at the $\tau \in (0,1)$ quantile with the k-th variable and $\varepsilon_{\tau i}$ is random residuals of the regression model at the τ-th quantile and i-th observation

2.5. Quantile Regression Parameter Estimation

In classical OLS regression, the parameters are estimated by minimizing the sum of squared residuals, whereas in quantile regression, they minimize the absolute sum of residuals, commonly known as the Least Absolute Deviation (LAD). The quantile weights are determined by τ, defined as τ if the residual value is greater than or equal to zero and 1-τ for residuals less than zero (Biswas et al., 2017). The estimated value of parameter β from the τ-th quantile regression can be obtained by Equation (9) and (10).

$$\hat{\beta}(\tau) = \min_{\beta}\left\{\tau \sum_{i=1;y>x}^{n} |\boldsymbol{y} - \boldsymbol{x^T\beta}| + (1-\tau)\sum_{i=1;y<x}^{n} |\boldsymbol{y} - \boldsymbol{x^T\beta}|\right\} \tag{9}$$

$$\hat{\beta}(\tau) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(\varepsilon_i) \tag{10}$$

for $$\rho_\tau(\varepsilon_i) = \begin{cases} \tau\varepsilon_i & \text{if} \quad \varepsilon_i \geq 0 \\ (\tau-1)\varepsilon_i & \text{if} \quad \varepsilon_i < 0 \end{cases} \tag{11}$$

where $\rho_\tau(\varepsilon)$ is referred to as the loss function, that is the multiplication between the residuals and the quantile weights of τ. The LAD method to obtain an estimate β of Equation (10) or (11) is done by minimizing the loss function. The solution cannot be obtained analytically but is obtained numerically using the simplex method, interior point method, or smoothing method. On the other hand, the estimation of quantile regression parameters with X variable being a binary category, can be estimated by solving as follows (Buhai, 2004).

$$(\hat{\alpha}_\tau, \hat{\beta}_\tau) = \underset{(\alpha,\beta)}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \delta D_i) \tag{12}$$

for $$\begin{cases} \hat{\alpha}_\tau = F_m^{-1}(\tau) \\ \hat{\delta}_\tau = G_n^{-1}(\tau) - F_m^{-1}(\tau) \end{cases} \tag{13}$$

where $\boldsymbol{G_n}$ and $\boldsymbol{F_m}$ represent the empirical distribution functions of the categorical and control variables, based on n and m observations respectively.

2.6. LASSO Regression

The Least Absolute Shrinkage and Selection Operator (LASSO) method was first introduced by Tibshirani in 1996. LASSO shrinks the regression coefficients of predictor variables that have a high correlation with the errors, to exactly zero or close to zero (Tibshirani, 1996). The general LASSO equation is formulated as (Zhao & Yu, 2006).

$$\boldsymbol{Y}^{**} = \boldsymbol{X}^{**}\boldsymbol{\beta}^{*} + \boldsymbol{\varepsilon}^{**} \quad (14)$$

with $\boldsymbol{Y}^{**}$ is response variable vector of size ($n \times 1$), $\boldsymbol{X}^{**}$ is matrix of predictor variables of size ($n \times (k+1)$), $\boldsymbol{\beta}^{*}$ is the vector of LASSO coefficients of size ($(k+1) \times 1$), and $\boldsymbol{\varepsilon}^{**}$ is error vector of size ($n \times 1$).

The estimation of LASSO coefficients uses quadratic programming with inequality constraints. LASSO estimation is obtained from Equation (15)

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \right\} \quad (15)$$

Under the condition that $\sum_{j=1}^{k} |\beta_j| \leq t$. The value of t is a tuning parameter that controls the shrinkage of the LASSO coefficient with $t \geq 0$. According to Tibshirani (1996), if $t < t_0$ with $t_0 = \sum_{j=1}^{k} |\hat{\beta}_j|$ it will cause the coefficient to shrink close to zero or exactly at zero, so that LASSO will have a role as variable selection. However, if $t > t_0$ then the LASSO coefficient estimator gives the same results as the least squares estimator method.

2.7. LASSO Regression Parameter Estimation Using CDA

Coordinate Descent Algorithm (CDA) is a model selection method where the algorithm can be modified and implemented into the LASSO Regression model (Hastie et al., 2016). Coordinate Descent performs well and quickly in solving the problem since each coordinate minimization can be done quickly and the relevant equations can thus be updated as we select variables (Friedman et al., 2007). The concept of the CDA algorithm in LASSO is optimizing the parameters separately and then optimizing the variables that are not optimal until they are optimal. Optimization is performed on a grid value λ, ranging from λ_{max} to λ_{min} as a tuning parameter in controlling the LASSO regression coefficients. The calculation steps using the CDA algorithm are as follows:

1. Standardize the data to have a mean value of 0 and a variance of 1.
2. Initialize all $\beta_j = 0$ and the loop is then run until convergence where the coefficient values are stable and do not change. Each coefficient is updated, and the soft-thresholding operator is applied.
3. Calculating LASSO simple least squares regression coefficients $\beta_j^* = \frac{1}{n}\sum_{i=1}^{n} x_{ij} r_{ij}$ with $r_{ij} = y_i - \sum_{p \neq j} x_{ip}\beta_p$. β_j^* is the LASSO regression parameter value.
4. Update the value of β_j^* by soft thresholding up to the optimum.

2.8. Ensemble Model

The fundamental concept behind the ensemble method is to combine several outputs from forecasting methods. Ensemble model can reduce the risk of overfitting thanks to the diversity of single model (Mohammed & Kora, 2023) and capture the uncertainty contained in each model This technique is prediction methods, especially in climate prediction. There are two main steps to building an ensemble. The first step is to create a membership of the ensemble and then combine the output of the ensemble members to produce a new ensemble output (Lusia & Suhartono, 2013). The two methods most commonly used in ensembles are averaging and stacking (Jafarzadeh et al., 2021; Lu et al., 2023).

1. Averaging. Using the averaging method, the output of the ensemble is obtained by averaging the ensemble member's output. Suppose *N* is the number of members in the ensemble, Equation 16 is obtained as follows:

$$y_{ensemble} = \frac{1}{N}\sum_{n=1}^{N} \hat{y}_N \, , n = 1,2, \ldots, N \quad (16)$$

2. Stacking. Stacking or stacked generalization is a generalized method of using a combination of higher-level models and lower-level models to achieve higher prediction accuracy. The global result of the ensemble can be calculated using the equation:

$$y_{ensemble} = \sum_{k=1}^{N} C_k \hat{y}_k \quad (17)$$

Breiman (1996), proposes to minimize the function G to be able to provide a better generalization of the model, which is

$$G = \sum_{t=1}^{n} [y_t - \sum_{k=1}^{N} C_k \hat{y}_k]^2 \quad (18)$$

by using the constraints $\sum_{k=1}^{N} C_k = 1$ and $0 \leq C_k \leq 1$

3. MATERIAL AND METHOD

3.1. Data and Variable

The data employed obtained from the 2018-2019 National Socio-Economic Survey (SUSENAS) conducted by the Central Statistics Agency. The research variables comprise response variables (y) and predictor variables (X). The response variable is per capita consumption in Surabaya City. The predictor variables are delineated in Table 2.

Table 2. Research Variable

Variable	Description
y	Per capita Consumption
X_1	Car Ownership
X_2	Refrigerator Ownership
X_3	Motorcycle Ownership
X_4	Computer Ownership
X_5	Home Phone Ownership
X_6	Ownership of Gas Cylinders of 5,5 kg or more
X_7	Air Conditioner (AC) Ownership
X_8	Water Heater Ownership
X_9	Gold Ownership
X_{10}	Boat Ownership
X_{11}	TV Ownership
X_{12}	Land Ownership
X_{13}	Fuel for Cooking: Electricity
X_{14}	Cooking Fuel: > 3 kg gas, 3 gas, town gas
X_{15}	Source of Drinking Water: Branded Bottled Water
X_{16}	Source of Drinking Water: Refill Water
X_{17}	Source of Drinking: Tap
X_{18}	Number of Households Members
X_{19}	Tenure Status of Occupied Residential Building: Owned
X_{20}	Tenure Status of Occupied Residential Building: Contract/Rent
X_{21}	Tenure Status of Occupied Residential Building: Rent-free
X_{22}	Tenure Status of Occupied Residential Building: Office
X_{23}	Log Floor Area per Capita
X_{24}	Number of households that have attended school
X_{25}	Number of households in which members have completed primary school
X_{26}	Number of households with members who have completed D1/D2/D3 education levels
X_{27}	Number of households with members who have completed S1/S2/S3 education levels
X_{28}	Number of households in which members have completed high school

X_{29}	Number of households with members who have completed junior high school
X_{30}	Number of households with members currently attending primary school.
X_{31}	Number of Households Currently attending junior high school
X_{32}	Number of households with members holding D1/D2/D3 qualifications
X_{33}	Number of households with members holding S1/S2/S3 education qualifications
X_{34}	Number of individuals currently attending senior high school
X_{35}	Number of households engaged in wholesale and retail trade, as well as the repair and maintenance of cars and motorcycles.
X_{36}	Number of Households Employed in Accommodation and Food and Beverage Services
X_{37}	Number of Households Working in Industry
X_{38}	Number of Households Working in Education
X_{39}	Number of Households Working in Transportation and Warehousing
X_{40}	Widest Floor Type: Marble/Granite/Ceramic, Parquet/Vinyl/Tapestry
X_{41}	Broadest Floor Type: Tile/Seal/Brasso
X_{42}	Broadest Floor Type: Wood/board, Cement/red brick
X_{43}	Widest Wall Type: Wall
X_{44}	Widest Wall Type: Wood/Board
X_{45}	Toilet Type: Private with Swan Neck Toilet Type
X_{46}	Toilet Type: Private with Slab Toilet Type
X_{47}	Number of Households with Disease
X_{48}	Number of Households with Health Coverage
X_{49}	Main Source of Lighting: PLN
X_{50}	The Widest Type of Roof is Concrete
X_{51}	The Widest Type of Roofing Tile
X_{52}	The Widest Roof Type is Asbestos
X_{53}	Fecal Landfill: Septic Tank
X_{54}	Place of Final Disposal of Feces: Hole in the Ground, Rice Field, River
X_{55}	Number of Working Households

3.2. Step of Analysis

The PMT analysis steps for estimating per capita consumption are as follows:

1. Preprocess the SUSENAS data by aggregating individual units into household units for analysis. Establish the response variable as household consumption per capita and include predictor variables such as asset ownership, education, and health.
2. Develop a regression model by employing four regression approaches, namely OLS regression, Robust regression, Quantile regression, and LASSO regression. The four methods are expected to overcome infringements of linear regression assumptions such as multicollinearity, autocorrelation and heteroscedasticity.
3. Create an ensemble model by combining the four regression approaches using an averaging method, as described in Equation 16.
4. Acquire the model for estimating household consumption expenditure.

4. RESULTS AND DISCUSSION

4.1. PMT Model

The construction of the PMT model to estimate per capita consumption utilizes four approaches: OLS, Robust, Quantile, and LASSO regression. The outcomes of the per capita consumption model estimation from each of these methods are presented below:

1. OLS Regression Model

$$\hat{y}_{OLS} = 13.615 + 0.258x_1 + 0.067x_2 + 0.051x_3 + 0.147x_4 + 0.041x_5 + 0.175x_6 + 0.123x_7 + 0.248x_8 + 0.093x_9 - 0.067x_{10} + 0.035x_{11} + 0.047x_{12} - 0.286x_{13} - 0.291x_{14} - 0.266x_{15} + 0.136x_{16} + 0.099x_{17} - 0.197x_{18} + 0.211x_{19} + 0.263x_{20} + 0.190x_{21} + 0.190x_{22} + 0.259x_{23} + 0.015x_{24} + 0.004x_{25} + 0.047x_{26} + 0.053x_{27} + 0.040x_{28} + 0.031x_{29} - 0.020x_{30} - 0.011x_{31} - 0.089x_{32} + 0.115x_{33} + 0.012x_{34} + 0.005x_{35} + 0.028x_{36} - 0.009x_{37} - 0.087x_{38} + 0.007x_{39} + 0.022x_{40} - 0.007x_{41} - 0.110x_{42} + 0.067x_{43} + 0.067x_{44} - 0.075x_{45} - 0.015x_{46} - 0.007x_{47} - 0.011x_{48} + 0.137x_{49} + 0.179x_{50} + 0.159x_{51} + 0.147x_{52} + 0.016x_{53} - 0.005x_{54} + 0.073x_{55}$$

2. Robust Regression Model

$$\hat{y}_{ROB} = 13.551 + 0.256x_1 + 0.076x_2 + 0.062x_3 + 0.137x_4 + 0.053x_5 + 0.168x_6 + 0.132x_7 + 0.205x_8 + 0.085x_9 - 0.034x_{10} + 0.049x_{11} + 0.037x_{12} - 0.286x_{13} - 0.313x_{14} + 0.298x_{15} + 0.180x_{16} + 0.136x_{17} - 0.194x_{18} + 0.165x_{19} + 0.201x_{20} + 0.126x_{21} + 0.132x_{22} + 0.243x_{23} + 0.017x_{24} + 0.002x_{25} + 0.051x_{26} + 0.040x_{27} + 0.038x_{28} + 0.018x_{29} - 0.019x_{30} - 0.005x_{31} - 0.096x_{32} + 0.097x_{33} + 0.028x_{34} + 0.013x_{35} + 0.035x_{36} - 0.002x_{37} - 0.062x_{38} + 0.020x_{39} + 0.000x_{40} - 0.086x_{41} - 0.122x_{42} + 0.070x_{43} + 0.086x_{44} - 0.102x_{45} - 0.040x_{46} - 0.006x_{47} - 0.010x_{48} + 0.248x_{49} + 0.182x_{50} + 0.155x_{51} + 0.150x_{52} + 0.036x_{53} + 0.036x_{54} + 0.067x_{55}$$

3. Quantile Regression Model

$$\hat{y}_{QUA} = 13.408 + 0.222x_1 + 0.060x_2 + 0.044x_3 + 0.153x_4 + 0.052x_5 + 0.173x_6 + 0.144x_7 + 0.203x_8 + 0.056x_9 - 0.005x_{10} + 0.042x_{11} + 0.014x_{12} - 0.286x_{13} - 0.307x_{14} + 0.274x_{15} + 0.142x_{16} + 0.077x_{17} - 0.193x_{18} + 0.218x_{19} + 0.226x_{20} + 0.160x_{21} + 0.163x_{22} + 0.259x_{23} + 0.012x_{24} + 0.005x_{25} + 0.029x_{26} + 0.040x_{27} + 0.021x_{28} + 0.010x_{29} - 0.025x_{30} + 0.000x_{31} - 0.095x_{32} + 0.127x_{33} + 0.043x_{34} + 0.023x_{35} + 0.056x_{36} - 0.013x_{37} - 0.062x_{38} + 0.022x_{39} + 0.085x_{40} - 0.022x_{41} - 0.042x_{42} + 0.095x_{43} + 0.152x_{44} - 0.112x_{45} - 0.076x_{46} - 0.005x_{47} - 0.006x_{48} + 0.268x_{49} + 0.211x_{50} + 0.172x_{51} + 0.174x_{52} + 0.038x_{53} + 0.012x_{54} + 0.069x_{55}$$

4. LASSO Regression Model

$$\hat{y}_{LAS} = 13.772 + 0.258x_1 + 0.060x_2 + 0.051x_3 + 0.146x_4 + 0.040x_5 + 0.174x_6 + 0.124x_7 + 0.248x_8 + 0.093x_9 - 0.066x_{10} + 0.035x_{11} + 0.048x_{12} - 0.286x_{13} - 0.291x_{14} + 0.219x_{15} + 0.089x_{16} + 0.052 - 0.196x_{18} + 0.183x_{19} + 0.235x_{20} + 0.161x_{21} + 0.161x_{22} + 0.258x_{23} + 0.014x_{24} + 0.004x_{25} + 0.047x_{26} + 0.052x_{27} + 0.040x_{28} + 0.031x_{29} - 0.020x_{30} - 0.011x_{31} - 0.087x_{32} + 0.115x_{33} + 0.012x_{34} + 0.005x_{35} + 0.028x_{36} - 0.009x_{37} - 0.086x_{38} + 0.007x_{39} + 0.024x_{40} - 0.076x_{41} - 0.010x_{42} + 0.051x_{43} + 0.065x_{44} - 0.073x_{45} - 0.012x_{46} - 0.007x_{47} - 0.011x_{48} + 0.140x_{49} + 0.095x_{50} + 0.075x_{51} + 0.063x_{52} + 0.015x_{53} - 0.005x_{54} + 0.073x_{55}$$

4.2. Estimation Results of Per Capita Consumption by Method Used

Upon acquiring the estimated model, the subsequent step involves conducting statistical tests to ascertain whether there exists an average difference between the estimated per capita consumption value from the formed model ($\hat{y}$) and the observed per capita consumption value (SUSENAS) (y). Figure 1 depicts overlapping intervals between the observed (y) and estimated ($\hat{y}$) values of the five models. This suggests that the average observed data (SUSENAS) aligns with the estimated results derived from the five models.

This result is corroborated by conducting a test for differences in the mean values between the five models and the observed values. Testing the difference in mean values with the following null hypothesis (H_0): there is no mean difference between the observed value (y) and the estimated result ($\hat{y}$) and alternative hypothesis (H_1): there is a mean difference between the observed value (y) and the estimated result ($\hat{y}$). Decision: Reject H_0 if the p-value < α=0.05.

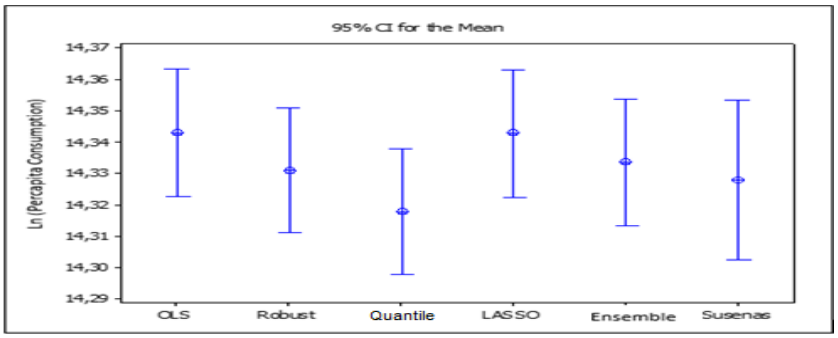

Figure 1. Interval Plot of 95% Median Value Estimation Results of PMT Model

Table 3 demonstrates that there is no discrepancy in the mean values of the estimated per capita consumption between the five models and the observed values. Alternatively, the estimation results obtained from the four models, namely OLS regression, Quantile regression, Robust regression, and LASSO regression, can be integrated into an Ensemble model. Figure 2 demonstrates that the distribution of observed data (y) and estimated values (ŷ) tend to align closely. Figure 3 illustrates that the observed variable, denoted as y, and the estimated variable, denoted as ŷ, exhibit a data distribution with no significant shift, as their curve shapes are nearly identical. Likewise, the ensemble model curve, derived from the averaged combination of four models, closely approximates the observed values (y). It is imperative to evaluate the new PMT model to ascertain whether its performance surpasses that of the previous PMT model. Evaluation can be carried out by assessing the estimation errors in comparison to the actual data, utilizing the MSE (Mean Square Error), MAE (Mean Absolute Error), and sMAPE (Symmetric Mean Absolute Percentage Error) criteria. The best model is identified as the one that exhibits the smallest values for MSE, MAE, and sMAPE.

Table 3. Hypothesis Testing of Mean y and $\hat{y}$ PMT Model

Hypothesis	*t-test*	*p-value*	Summary
$\hat{y}_{OLS}$ vs y	-1.73	0.084	The calculated mean of the OLS model does not demonstrate a statistically significant difference from the observed values.
$\hat{y}_{Robust}$ vs y	-0.59	0.557	The calculated mean of the Robust model does not show a statistically significant difference from the observed values.
$\hat{y}_{Quantil}$ vs y	-0.50	0.619	The calculated mean of the Quantile model does not exhibit a statistically significant difference from the observed values.
$\hat{y}_{Lasso}$ vs y	-1.00	0.315	The calculated mean of the LASSO model does not display a statistically significant difference from the observed values.
$\hat{y}_{Ensemble}$ vs y	-0.96	0.339	The calculated mean of the ensemble does not exhibit a statistically significant difference from the observed values.

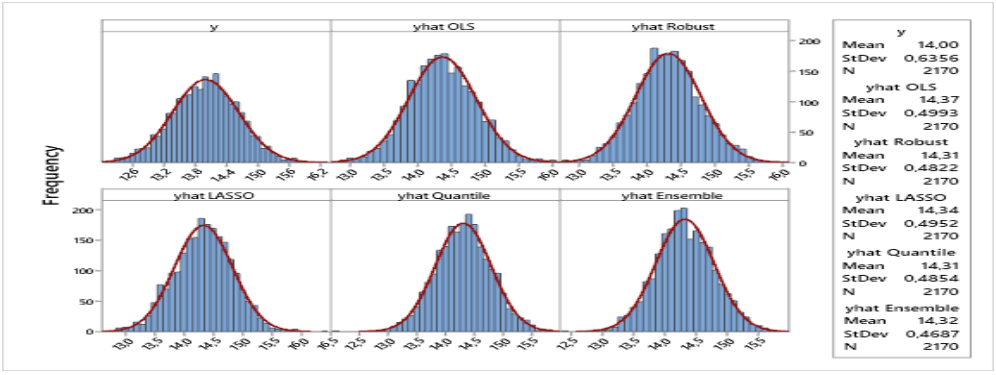

Figure 2. Comparison of Distribution Spread of y and ŷ PMT Data

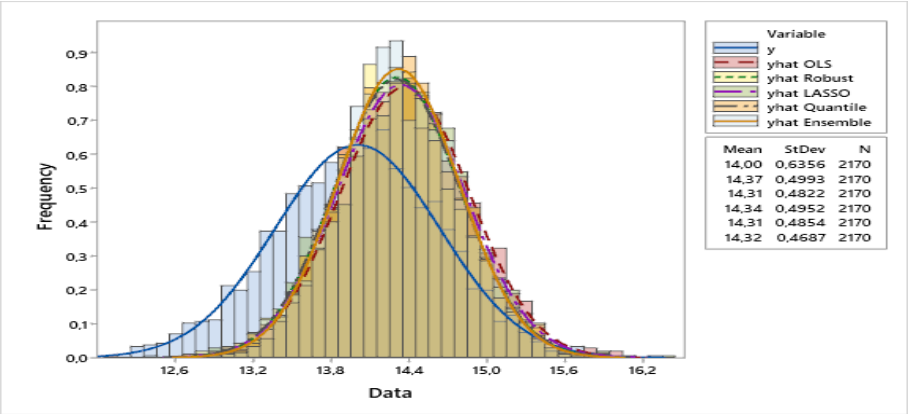

Figure 3. Distribution of PMT Model Estimation Results

Table 4. Model Evaluation

Model Evaluation	Methods				
	OLS	Robust	Quantile	LASSO	Ensemble
MSE	0.145	0.146	0.147	0.144	0.145
MAE	0.286	0.284	0.284	0.285	0.284
sMAPE	0.996%	0.989%	0.986%	0.991%	0.988%

Table 4 indicates that the MSE of LASSO regression yields the smallest error value. However, in terms of the MAE criterion, Robust regression, Quantile regression, and the Ensemble model display the smallest error values. Regarding the sMAPE criterion, Quantile regression exhibits the smallest error value. Based on the model evaluation, none of the models exhibits the highest accuracy in estimating per capita consumption within this research. Generally, there is no absolute best model in the proxy mean test (Houssou et al., 2007; Kidd & Wylde, 2011), thus it is necessary to combine models to capture all the information contained in a model. The application of the ensemble method is not only as a comparison between models, but also as a means to capture the uncertainty contained in each model (Tran et al., 2020; Vrugt & Robinson, 2007).

4.3. Determination of Household Groups by Per Capita Consumption

The determination of a household's decile placement relies on the per capita expenditure distribution derived from SUSENAS data, as illustrated in Table 5. This decile classification is established using SUSENAS data, which divides the data range into ten groups. Deciles 1 to 4 encompass the spectrum of families within the low-income family's category, each having a per capita expenditure limit of less than or equal to IDR 1,267,774.

Table 5. Clustering of Decile Intervals

Decile	Mean	Minimum	Maximum	Decile	Mean	Minimum	Maximum
1	607.315	327.345	726.499	6	1.612.409	1.504.389	1.725.220
2	811.955	726.678	887.798	7	1.860.208	1.726.857	2.007.693
3	980.394	890.706	1.070.339	8	2.201.284	2.008.856	2.408.431
4	1.164.783	1.071.472	1.267.774	9	2.763.869	2.409.232	3.273.881
5	1.376.767	1.268.020	1.503.450	10	6.016.971	3.279.403	80.906.952

5. CONCLUSION

The conclusion drawn from the results and discussion is that the optimal approach for estimating per capita consumption involves employing an ensemble model, which

comprises a combination of OLS regression, robust regression, quantile regression, and LASSO regression. The category of families falling within the low-income family's classification comprises those with a per capita expenditure limit of less than or equal to IDR 1,267,774. In other words, this range of families, characterized by these expenditure limits, can serve as a reference for the Surabaya City Government when devising assistance programs or social initiatives for the impoverished population of Surabaya City. Future research suggestions may involve the exploration of additional variables or factors that could serve as indicators for categorizing low-income families in Surabaya City, thereby enhancing the precision and accuracy of the results.

ACKNOWLEDGMENT

We would like to express our gratitude to the Central Bureau of Statistics for granting us the opportunity to utilize their data in the course of this research.

REFERENCES

Amédée-Manesme, C. O., Faye, B., & Fur, E. Le. (2020). Heterogeneity and fine Wine Prices: Application of the Quantile Regression Approach. *Applied Economics*, *52*(26), 2821-2840.

Biswas, J., Kulkarni, H., & Das, K. (2017). Quantile Regression in Biostatistics. *Biostat Biometrics Open Access Journal*, *2*(5), 102–105.

Blanc, S. M. & Setzer, T. (2016). When to Choose the Simple Average in Forecast Combination. *Journal of Business Research*, *69*(10), 3951–3962.

BPS. (2021). *Provinsi Jawa Timur Dalam Angka 2021*.

Breiman, L. (1996). Stacked Regressions. *Machine Learning*, *24*(1), 49–64.

Buhai, S. (2004). Quantile Regression: Overview and Selected Applications. *Journal of AD Astra*, *4*(4), 1–16.

Chen, C. (2002). Robust Regression and Outlier Detection with the ROBUSTREG Procedure. *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*.

Friedman, J., Hastie, T., Hofling, H., & Tibshirani, R. (2007). Pathwise Coordinate Optimization. *The Annual of Applied Statistics*, *1*(2), 302–332.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)* (2nd ed.). Springer.

Hong, H. G., Christiani, D. C., & Li, Y. (2019). Quantile Regression for Survival Data in Modern Cancer Research: Expanding Statistical Tools for Precision Medicine. *Precision Clinical Medicine*.

Houssou, N., Zeller, M., Alcaraz V., G., Schwarze, S., & Johannsen, J. (2007). Proxy Means Tests for Targeting the Poorest Households -- Applications to Uganda. *106th Seminar of the EAAE Pro-Poor Development in Low Income Countries: Food, Agriculture, Trade, and Environment 2*, *October*.

Jafarzadeh, A., Pourreza-Bilondi, M., Akbarpour, A., Khashei-Siuki, A., & Samadi, S. (2021). Application of Multi-Model Ensemble Averaging Techniques for Groundwater Simulation: Synthetic and Real-world Case Studies. *Journal of Hydroinformatics*, *23*(6), 1271–1289.

Khan, D. M., Ali, M., Ahmad, Z., Manzoor, S., & Hussain, S. (2021). A New Efficient Redescending M-Estimator for Robust Fitting of Linear Regression Models in the Presence of Outliers. *Mathematical Problems in Engineering*, *2021*.

Kidd, S. & Wylde, E. (2011). Targeting the Poorest: An Assessment of the Proxy Means Test Methodology. *Technical Report Australian Government* (Issue September).

Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill Companies, Inc.

Lu, M., Hou, Q., Qin, S., Zhou, L., Hua, D., Wang, X., & Cheng, L. (2023). A Stacking Ensemble Model of Various Machine Learning Models for Daily Runoff Forecasting. *Water (Switzerland)*, *15*(7).

Lusia, D. A. & Suhartono. (2013). Ensemble Method Based on Two Level ARIMAX- FFNN for Rainfall Forecasting in Indonesia. *International Journal of Science and Research*, *2*(2), 144–149.

Mohammed, A., & Kora, R. (2023). A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 757–774.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (1992). *Introduction to Linear Regression Analysis* (2nd ed.). John Wiley & Sons, Inc.

Peng, X., Zheng, W., Zhang, D., Liu, Y., Lu, D., & Lin, L. (2017). A Novel Probabilistic Wind Speed Forecasting Based on Combination of the Adaptive Ensemble of On-line Sequential ORELM (Outlier Robust Extreme Learning Machine) and TVMCF (Time-varying Mixture Copula function). *Energy Conversion and Management*, *138*, 587–602.

Pratiwi, H., Susanti, Y., & Handajani, S. S. (2018). A Robust Regression by Using Huber Estimator and Tukey Bisquare Estimator for Predicting Availability of Corn in Karanganyar Regency, Indonesia. *Indonesian Journal of Applied Statistics*, *1*(1), 37.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tran, K., Neiswanger, W., Yoon, J., Zhang, Q., Xing, E., & Ulissi, Z. W. (2020). Methods for Comparing Uncertainty Quantifications for Material Property Predictions. *Machine Learning: Science and Technology*, *1*(2).

Vrugt, J. A. & Robinson, B. A. (2007). Treatment of Uncertainty using Ensemble Methods: Comparison of Sequential Data Assimilation and Bayesian Model Averaging. *Water Resources Research*, *43*(1), 1–15.

Wang, S., Zhang, N., Wu, L., & Wang, Y. (2016). Wind Speed Forecasting Based on the Hybrid Ensemble Empirical Mode Decomposition and GA-BP Neural Network Method. *Renewable Energy*, *94*, 629–636.

Zhang, G. P. (2003). Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing*, *50*, 159–175.

Zhao, P., & Yu, B. (2006). On Model Selection Consistency of LASSO. *Journal of Machine Learning Research*, *7*, 2541–2563.