PENINGKATAN KUALITAS STATISTIK RESMI PRODUKTIVITAS PADI MELALUI IMPUTASI DATA NON-RESPONS MENGGUNAKAN MODEL ADITIF GEOSPASIAL

Muhlis Ardiansyah[1,2]

[1]BPS Kabupaten Kotawaringin Timur, Kalimantan Tengah, Indonesia
[2] Departemen Statistika, IPB University, Kampus IPB Dramaga, Bogor, Indonesia
e-mail: muhli@bps.go.id; muhlisardiansyah@apps.ipb.ac.id

Abstrak

Penelitian ini dilatarbelakangi oleh masalah non-respons pada Survei Ubinan yang diselenggarakan oleh Badan Pusat Statistik (BPS) sebagai penyedia statistik resmi. BPS memiliki visi sebagai penyedia data statistik berkualitas untuk Indonesia maju. Penanganan non-respons sangat penting untuk mendukung visi tersebut karena non-respons berpotensi menyebabkan beberapa karakteristik sampel menjadi tidak terwakili. Penelitian ini mengusulkan teknik imputasi data non-respon melalui pemodelan statistik. Model yang diusulkan adalah model aditif dengan penambahan fungsi pemulusan geospasial *thin plate regression splines* (TP) dan *Gaussian process* (GP). Pemilihan model terbaik berdasarkan *Mean Squared Error of Prediction* (MSEP) terkecil dari 1000 iterasi. Kemudian dibandingkan rata-rata produktivitas padi antara teknik mengabaikan non-respons (*listwise deletion*) dan imputasi melalui tiga skenario data non-respons. Hasil penelitian menunjukkan bahwa model dengan penambahan fungsi pemulus GP memberikan kinerja terbaik dengan MSEP tekecil. Hasil lainnya menunjukkan bahwa metode imputasi data non-respons lebih baik dibandingkan dengan mengabaikan non-respons. BPS dapat mempertimbangkan metode imputasi untuk meningkatkan kualitas data statistik resmi produktivitas padi.

Kata Kunci: imputasi, geospasial, non-respons, statistik resmi

Abstract

This study is motivated by the non-response problem in the Crop Cutting Survey conducted by the BPS-Statistics Indonesia as the official statistics provider. BPS has a vision of providing quality statistical data for advanced Indonesia. Handling non-response is essential to supporting this vision because non-response can potentially cause some sample characteristics to be unrepresented. This study proposed a non-response data imputation technique through statistical modeling. The proposed model was an additive model with the addition of geospatial smoothing functions of thin plate regression splines (TP) and Gaussian process (GP). Selection of the best model based on the smallest MSEP of 1000 iterations. Then we compared the average rice productivity between listwise deletion and imputation techniques through three scenarios of non-response data. The results showed that the model with the addition of the GP smoothing function gave the best performance with the smallest MSEP. The other results showed that the imputation method of non-response data is better than ignoring non-response. BPS can consider the imputation method to improve the quality of official statistics on rice productivity.

***Keyword**: imputation, geospatial, non-response, official statistics*

PENDAHULUAN

Badan Pusat Statistik (BPS) sebagai lembaga penyedia statistik resmi (*official statistics*) melakukan penghitungan produktivitas tanaman padi (*Oryza sativa*) melalui Survei Ubinan. Data resmi produktivitas padi dibutuhkan untuk mendapatkan informasi tentang capaian target kedua pembangunan berkelanjutan (*Sustainable Development Goals* - SDGs). Data tersebut dihitung dengan mengambil sampel padi yang siap panen pada petak sawah maupun bukan sawah kemudian mengambil secara acak satu plot padi berukuran $2.5 \times 2.5\, m^2$ untuk dipanen, dibersihkan, dan ditimbang. Salah satu permasalahan yang dihadapi pada pelaksanaan survei ini adalah terdapat non-respons yang cukup banyak (Ardiansyah *et al.,* 2021).

Data non-respons pada Survei Ubinan disebabkan oleh beberapa hal. Pertama, petugas kadang-kadang tidak berhasil mengambil sampel plot tanaman padi karena sudah dipanen oleh petani terpilih. Kedua, kebijakan pembatasan sosial oleh pemerintah. Ketiga, perubahan cara pemanenan di beberapa petani dari cara tebas ke mesin *combine harvester* sehingga menyebabkan petugas tidak dapat mengambil sampel (Ardiansyah *et al.,* 2020).

Data non-respons menyebabkan berkurangnya keterwakilan sampel sehingga sebaran sampel plot ubinan menjadi tidak merata. Jika sampel yang dipilih berada pada kelompok dengan produktivitas padi rendah maka hasil estimasi akan lebih rendah dari nilai sebenarnya (*underestimate*) dan begitupun sebaliknya. Guna mendukung visi BPS sebagai penyedia data statistik berkualitas untuk Indonesia maju (*provider of qualified statistical data for advanced Indonesia*), penanganan non-respons menjadi penting untuk dilakukan guna meningkatkan kualitas data produktivitas padi.

Terdapat tiga kerugian yang disebabkan adanya data non-respons pada suatu survei. Pertama, berkurangnya kekuatan statistik akibat ukuran sampel yang berkurang. Hal ini berpotensi meningkatkan risiko kesalahan jenis II (*Error type II*) yang merupakan kesalahan akibat menerima H_0 padahal H_1 benar. Kedua, data non-respons berpotensi menimbulkan bias pendugaan. Ketiga, dugaan galat baku menjadi *over* atau *underestimate* (Curley *et al.,* 2019).

Penelitian ini merupakan kelanjutan dari penelitian sebelumnya. Pada penelitian sebelumnya, Ardiansyah & Tofri (2019) mengusulkan metode wawancara pascapanen berdasarkan pengakuan petani untuk mengatasi permasalahan non-respons pada Survei Ubinan. Hasilnya, metode wawancara pasca panen belum cukup bukti untuk menggantikan metode pengukuran plot ubinan karena memberikan hasil produktivitas padi yang lebih rendah (*underestimate*). Oleh karena itu, penelitian ini mengusulkan cara lain yaitu teknik imputasi data non-respons melalui pemodelan statistika.

Secara garis besar, terdapat tiga cara untuk menangani data non-respons pada suatu survei. Pertama, petugas melakukan kunjungan ulang (*revisit*) untuk melengkapi data. Cara ini tidak mungkin dilakukan pada kasus Survei Ubinan karena sampel terpilih sudah terlanjur dipanen oleh petani sehingga tidak mungkin dilakukan pengukuran berat gabah. Cara kedua adalah dengan menghapus atau mengabaikannya (*listwise deletion*) yaitu pendugaan parameter hanya dilakukan dengan menggunakan data respon yang berhasil dikumpulkan dan mengabaikan data non-respons. Cara ketiga adalah dengan mengimputasinya (*imputation*) yaitu menduga data non-respons dengan metode tertentu.

Chhabra *et al.,* (2019) membagi cara imputasi data non-respons menjadi dua. Pertama, cara tradisional seperti imputasi rata-rata, imputasi acak sesuai kelasnya, imputasi hot-deck berurutan, dan imputasi hot-deck hirarki. Kedua, metode modern seperti metode imputasi berbasis regresi, imputasi berganda, dan imputasi menggunakan pembelajaran mesin.

Metode imputasi yang dikembangkan pada penelitian ini adalah metode imputasi berbasis regresi dengan memanfaatkan peubah lainnya. Pada Survei Ubinan, kita dihadapkan pada peubah respon yang sulit diukur dan peubah prediktor yang mudah diperoleh. Peubah prediktor tersebut akan dimanfaatkan untuk mengimputasi data non-respons.

Akurasi dan presisi imputasi data non-respons berbasis regresi dapat ditingkatkan dengan mempertimbangkan berbagai model statistika seperti *Generalized Linear Model* (GLM) (Nelder & Wedderburn, 1972), *Generalized Additive Models* (GAM) (Hastie & Tibshirani, 1986), *Generalized Linear Mixed Models* (GLMM) (Breslow & Clayton, 1993), dan *Generalized Additive Mixed Models* (GAMM) (Lin & Zhang, 1999). Seiring perkembangan teknologi, model dengan pendekatan berbasis data (*data-driven*) untuk mengeksploitasi pola nonlinier dapat ditempuh dengan metode gabungan antara parametrik dan nonparametrik. Pada penelitian ini akan ditambahkan fungsi pemulus geospasial pada model aditif yaitu *thin plate regression splines* (TP) dan *Gaussian process smooths* (GP). Model imputasi yang diajukan ada tiga, yaitu model campuran linier tanpa penambahan fungsi pemulus geospasial (GLMM), model campuran linier dengan penambahan fungsi pemulus TP (GLMM+TP), dan model campuran linier dengan penambahan fungsi pemulus GP (GLMM+GP). Model yang diajukan dipilih karena memiliki fleksibilitas sehingga diduga mampu memberikan akurasi imputasi yang lebih baik dibanding model tanpa penambahan fungsi pemulus geospasial.

Penelitian ini bertujuan untuk memilih metode imputasi terbaik antara model GLMM, GLMM+TP, dan GLMM+GP. Tujuan kedua adalah mengevaluasi kinerja metode imputasi dibandingkan metode mengabaikan atau menghapus data non-respons (*listwise deletion*). Imputasi dilakukan dengan mengganti data non-respons dengan nilai dugaan. Setelah semua data hilang diganti, dataset dianalisis menggunakan teknik standar sebagaimana pendugaan pada data yang lengkap.

METODE

Langkah pertama adalah mengidentifikasi apakah ada autokorelasi spasial pada data produktivitas padi menggunakan indeks Moran. Indeks Moran adalah ukuran autokorelasi spasial yang diusulkan pertama kali oleh Patrick Alfred Pierce Moran pada tahun 1950. Indeks Moran diformulasikan sebagai berikut:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i=1}^{n}(y_i-\bar{y})^2}, \quad \ldots\ldots\ldots\ldots (1)$$

dengan

y_i = nilai amatan lokasi ke$-i$ ($i = 1,2,\ldots,n$),

y_j = nilai amatan lokasi ke$-j$ ($j = 1,2,\ldots,n$),

$\bar{y}$ = rataan nilai amatan dari seluruh lokasi,

w_{ij} = pembobot kedekatan antara lokasi ke$-i$ dengan lokasi ke$-j$.

Pembobot yang dipilih adalah invers jarak. Informasi mengenai lokasi dilihat dari titik longitude dan latitude. Titik longitude dan latitude terlebih dahulu dinormalisasi. Informasi inilah yang digunakan untuk menghitung jarak antar titik sampel. Kekuatan ketergantungan spasial akan menurun sesuai dengan semakin jauhnya jarak. Jarak ditentukan dengan jarak Euclid. Uji signifikansi Indeks Moran perlu dilakukan untuk melihat nyata tidaknya autokorelasi spasial. Hipotesis untuk Indeks Moran adalah:

H_0 : $I = 0$ (Tidak terdapat autokorelasi spasial)

H_1 : $I > 0$ (Terdapat autokorelasi spasial.)

Statistik uji: $Z(I) = \frac{I-E(I)}{\sqrt{Var\ (I)}} \sim N(0,1)$ dengan nilai harapan Indeks Moran pada asumsi normal dan acak adalah $E(I) = -\frac{1}{n-1}$; $Var\ (I) = \frac{n^2S_1 - nS_2 + 3S_0^2}{(n^2-1)S_0^2} - \left[\frac{1}{n-1}\right]^2$; $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$; $S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij} + w_{ji}\right)^2$; $S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2$. Kriteria ujinya adalah tolak H_0 jika $|Z(I)| > Z_\alpha$ (Djuraidah 2020).

Model Aditif Terampat (*Generalized Additive Models*/ GAM)

Model aditif terampat merupakan perpaduan antara peubah prediktor parametrik dan nonparametrik atau yang sering disebut model semiparametrik dengan peubah respon termasuk ke dalam keluarga eksponensial. GAM pertama kali diperkenalkan oleh Hastie dan Tibshirani pada tahun 1986 yang menganggap bahwa mean dari peubah respon yang termasuk dalam keluarga eksponensial tergantung pada prediktor berupa fungsi aditif melalui fungsi hubung (*link function*).

Misalkan $\boldsymbol{\mathcal{X}}^{[j]}$ menunjukkan matriks model dan $\boldsymbol{\mathcal{S}}^{[j]}$ menunjukkan matriks penalti untuk f_j. Kemudian digabungkan antara **A** dan $\boldsymbol{\mathcal{X}}^{[j]}$ berdasarkan kolom, untuk membuat model matriks $\mathbf{X} = \left(\mathbf{A}: \boldsymbol{\mathcal{X}}^{[1]}: \boldsymbol{\mathcal{X}}^{[2]}: \ldots\right)$. $\boldsymbol{\beta}$ adalah vektor koefisien model yang berisi γ dan vektor koefisien suku halus individu. Penalti pemulusan total untuk model dapat ditulis sebagai $\sum_j \lambda_j \boldsymbol{\beta}^{\mathrm{T}} \mathbf{S}_j \boldsymbol{\beta}$, dimana λ_j adalah parameter pemulus dan $\mathbf{S}_j$ hanyalah $\boldsymbol{\mathcal{S}}_j$ yang disematkan (*embedded*) sebagai blok diagonal dalam matriks yang mana selain elemen diagonal berisi nol, sehingga $\lambda_j \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta}$ adalah penalti untuk f_j. Model aditif secara umum memiliki bentuk: $g(\mu_i) = \mathbf{A}_i \gamma + \sum_j f_j(x_{ji})$, $y_i \sim \mathrm{EF}(\mu_i, \phi)$. Model aditif dapat dipandang sebagai model linier dengan parameter yang berlebih $g(\mu_i) = \mathbf{X}_i \boldsymbol{\beta}$, $y_i \sim \mathrm{EF}(\mu_i, \phi)$ dan parameternya diduga dengan memaksimalkan $l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2\phi} \sum_j \lambda_j \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta}$. Parameter λ_j mengontrol pertukaran antara kecocokan dan kehalusan model (Wood, 2017).

1. Pendugaan parameter $\boldsymbol{\beta}$ jika $\boldsymbol{\lambda}$ diketahui dapat dilakukan melalui *Penalized Iteratively Reweighted Least Squares* (PIRLS) dengan Langkah sebagai berikut (Wood, 2017):
2. Inisialisasi $\hat{\mu}_i = y_i + \delta_i$ dan $\hat{\eta}_i = g(\hat{\mu}_i)$ dimana δ_i biasanya nol, tetapi mungkin konstanta kecil yang memastikan bahwa $\hat{\eta}_i$ adalah berhingga. Ulangi dua langkah berikutnya hingga konvergen.
3. Hitung data semu $z_i = \frac{g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)}{\alpha(\hat{\mu}_i)} + \hat{\eta}_i$, dan bobot iteratif $w_i = \frac{\alpha(\hat{\mu}_i)}{\{g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)\}}$.
4. Temukan $\widehat{\boldsymbol{\beta}}$ untuk meminimalkan kuadrat terkecil terboboti. $\|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|_W^2 + \sum_j \lambda_j \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta}$ dengan $\|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|_W^2 = (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})$ dan $V(\mu)$ adalah fungsi varians yang ditentukan oleh sebaran keluarga eksponensial dan $\alpha(\mu_i) = \left[1 + (y_i - \mu_i)\left\{\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)}\right\}\right]$ atau dapat menggunakan pendekatan 'Fisher scoring' dimana Hessian dari kemungkinan log diganti dengan nilai harapannya, sesuai dengan pengaturan $\alpha(\mu_i) = 1$.
5. Perbarui $\hat{\eta} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ dan $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

Jika mengikuti hubungan antara GAM dan model campuran linier terampat (*Generalized Linear Mixed Model*-GLMM) dan mengidentifikasi $\lambda_j \mathbf{S}_j / \phi$ dengan matriks pengaruh acak, maka penduga REML yang sesuai untuk parameter skala adalah $\hat{\phi} = \frac{\|\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_W^2}{n - \tau}$ dengan $\tau = \mathrm{tr}\{(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\}$, dan $\mathbf{S}_\lambda = \sum_j \lambda_j \mathbf{S}_j$. τ dapat ditafsirkan sebagai derajat kebebasan efektif, dengan $\mathbf{F} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}$ adalah penghitungan bobot. Derajat kebebasan efektif untuk setiap suku mulus diperoleh dengan menjumlahkan nilai F_{ii} yang sesuai dengan koefisien β_i dari suku mulus.

Fungsi mulus dapat dipandang sebagai pengaruh acak dalam model campuran linier terampat yang parameternya dapat diduga menggunakan metode kemungkinan marginal. Kesulitan utama dengan cara kerja ini adalah bahwa fungsi mulus harus diatur dalam bentuk yang sesuai dengan struktur pengaruh acak. Misalnya, untuk setiap smooth dengan parameter smoothing tunggal yang dapat diekspresikan dalam bentuk $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$ dimana $\boldsymbol{\beta}$ diperlakukan sebagai vektor pengaruh tetap (tidak dihukum) dan $\mathbf{b} \sim N(\mathbf{0}, \mathbf{I}\sigma_b^2)$ diperlakukan sebagai vektor pengaruh acak (Wood *et al.*, 2013). Untuk pengaruh acak lebih dari satu, konstruksi matriks dirancang sedemikian rupa sehingga partisi sederhana dari kolom-kolom matriks model smooth dan vektor parameternya, memungkinkan smooths untuk direpresentasikan dalam bentuk $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{b}_1 + \mathbf{Z}_2\mathbf{b}_2 + \ldots$, dimana $\boldsymbol{\beta}$ merupakan vektor parameter tetap, dan $\mathbf{b}_j$ adalah vektor independen dari pengaruh acak $\mathbf{b}_j \sim N\left(\mathbf{0}, \mathbf{I}\sigma_{b_j}^2\right)$.

Model Aditif Geospasial

Model aditif geospasial adalah model aditif dengan menambahkan fungsi pemulus geospasial seperti *thin plate regression splines* (GLMM+TP) dan *Gaussian process* (GLMM+GP). Secara umum, model *additive* memiliki struktur sebagai berikut $g(\mu_i) = \boldsymbol{X}_i^* \theta + f_1(x_{1i}) + \cdots$ dengan $\mu_i = E(Y_i)$ dan Y_i memiliki sebaran yang termasuk ke dalam keluarga eksponensial.

Pertama, GLMM+TP adalah model aditif yang menggabungkan antara model

GLMM dan model *additive* dengan menambahkan fungsi geo-spasial *thin plate regression splines* (TP). Misalkan $\boldsymbol{S}_i = (s_{i1}, s_{i2})^T$ menunjukkan titik koordinat longitude dan latitude pada lokasi ke-i, $i = 1, \dots, n$. Misalkan $\boldsymbol{Y}_i$ adalah peubah respon dan $\boldsymbol{X}_i = (X_{i1}, \dots, X_{i2})^T$ adalah peubah prediktor pada lokasi $\boldsymbol{S}_i$. Kita asumsikan kepekatan peluang $(Y|\mathbf{x}, \mathbf{d}, \mathbf{s})$ termasuk anggota sebaran keluarga eksponensial dengan $\mu(\mathbf{x}, \mathbf{d}, \mathbf{s})$ dimodelkan dengan fungsi hubung $g(.)$ dalam bentuk aditif: $g\{\mu(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{s})\} = \sum_{k=0}^{p-1} \beta_k x_k + d_i u + \alpha(\boldsymbol{s})$ dengan u adalah pengaruh acak area, dan $\alpha(.)$ adalah fungsi mulus bivariate TP. Maka model GLMM+TP dapat diformulasikan dalam bentuk matriks:

$$\boldsymbol{g}\big(\boldsymbol{E}(\boldsymbol{Y})\big) = \boldsymbol{X\beta} + \boldsymbol{Du} + \boldsymbol{Zv}, \dots\dots\dots (2)$$

dengan $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})^T$ adalah parameter yang tidak diketahui, $\mathbf{X} = [\boldsymbol{X}_1 | \boldsymbol{X}_2]$; $\boldsymbol{X}_1 = [1, \boldsymbol{x}_{ij}^{\mathrm{T}}]_{1 \le i \le n}$; $\boldsymbol{X}_2 = [\boldsymbol{s}_{ij}^{\mathrm{T}}]_{1 \le i \le n}$, $\mathrm{E}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{v}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$, $\mathrm{Cov}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{v}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\sigma_u^2 \boldsymbol{I}_m & 0 & 0\\ 0 & \sigma_v^2 \boldsymbol{I}_{\boldsymbol{K_s}} & 0\\ 0 & 0 & \sigma_e^2 \boldsymbol{I}_n\end{bmatrix}$, **D** adalah matrik dengan pengaruh acak yang teramati berukuran $n \times m$ dengan m adalah banyaknya sub-populasi. Sedangkan $\boldsymbol{u}$ adalah vektor pengaruh acak yang tidak diketahui berukuran $m \times 1$, $\boldsymbol{u} \sim_{\text{iid}} \mathrm{N}(\mathbf{0}, \sigma_u^2 \boldsymbol{I}_m)$. $\boldsymbol{Z}$ adalah matriks berukuran $n \times K_s$ dari bivariate TP berdasarkan fungsi: $\boldsymbol{Z} = [C(s_i - \kappa_k^s)]_{1 \le i \le n,\, 1 \le k \le K_s}$, dan $\mathrm{C}(\boldsymbol{s}) = \|\boldsymbol{s}\|^2 \log\|\boldsymbol{s}\|$ dan $\kappa_1^s, \dots, \kappa_{K_s}^s$ adalah simpul (knots). $\boldsymbol{v}$ adalah vektor koefisien bivariate TP berukuran $K_s \times 1$, $\boldsymbol{v} \sim_{\text{iid}} \mathrm{N}(\mathbf{0}, \sigma_v^2 \boldsymbol{I}_{\boldsymbol{K_s}})$. Sedangkan $\boldsymbol{e}$ adalah galat acak, $\boldsymbol{e} \sim_{\text{iid}} \mathrm{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_n)$.

Kedua, GLMM+GP adalah model aditif yang menggabungkan antara model GLMM dan model *additive* dengan menambahkan fungsi geo-spasial *Gaussian Process* (GP). Misalkan $\boldsymbol{S}_i = (s_{i1}, s_{i2})^T$ menunjukkan titik koordinat longitude dan latitude pada lokasi ke-i, $i = 1, \dots, n$. Misalkan $\boldsymbol{Y}_i$ adalah peubah respon dan $\boldsymbol{X}_i = (X_{i1}, \dots, X_{i2})^T$ adalah peubah prediktor pada lokasi $\boldsymbol{S}_i$. Kita asumsikan kepekatan peluang $(Y|\mathbf{x}, \mathbf{d}, \mathbf{s})$ termasuk anggota sebaran keluarga eksponensial dengan $\mu(\mathbf{x}, \mathbf{d}, \mathbf{s})$ dimodelkan dengan fungsi hubung $g(.)$ dalam bentuk aditif. Jika $\mathrm{var}(Y|\mathbf{X} = \mathbf{x}, \mathbf{S} = \mathbf{s}) = \sigma^2 V\{\mu(\mathbf{x}, \mathbf{s})\}$, maka pendugaan nilai tengah (*mean*) dapat diperoleh dengan mengganti fungsi log-likelihood bersyarat $\big(\log\{f_{Y|\mathbf{X},\mathbf{S}}(y|\boldsymbol{x}, \boldsymbol{s})\}\big)$ dengan fungsi quasi-likelihood $l(\vartheta, y)$, yang memenuhi $\nabla_\vartheta l(\vartheta, y) = \frac{y - \vartheta}{\sigma^2 V(\vartheta)}$. Pendugaan ini menggunakan pendekatan quasi-likelihood nonparametrik (Yu *et al* 2019).

Misalkan $\alpha(.)$ adalah fungsi *additive Gaussian Process*, maka fungsi α adalah penjumlahan dari k fungsi regresi yang dikontrol oleh himpunan parameter $\emptyset = (\emptyset_1, \dots, \emptyset_k)^T$ atau $\alpha(s_i) = \emptyset_1 f_1(s_1) + \dots + \emptyset_k f_k(s_k)$. Maka GLMM+GP dapat diformulasikan mirip dengan GLMM+TP dalam bentuk matriks $\boldsymbol{g}\big(\boldsymbol{E}(\boldsymbol{Y})\big) = \boldsymbol{X\beta} + \boldsymbol{Du} + \boldsymbol{P\vartheta}$. *Gaussian process* (GP) menyediakan prior yang fleksibel untuk setiap fungsi komponen dalam $\{f_l, l = 1, \dots, k\}$ dimana $f_l \sim GP(0, c_l)$ dengan $c_l(s, s') = \exp\left\{-\sum_{j=1}^{p} K_{ij}\left(s_j - s_j'\right)^2\right\}$ dengan hyperprior Gamma yang ditetapkan ke parameter inverse-bandwidth K_{ij} memastikan pendugaan optimal dari fungsi regresi isotropic. Pendugaan parameter dari model ini dapat melihat Vo & Pati (2017).

Metode Evaluasi

Gugus data dibagi menjadi data training dan data testing dengan perbandingan 70:30 persen. Rasio perbandingan antara data training dan testing ditentukan oleh peneliti dengan asumsi terjadi non-respon sebanyak 30 persen. Model dibangun berdasarkan data training yang masing-masing gugus data diulang sebanyak 1000 kali. Kemudian dilakukan pendugaan berat gabah kg per $2.5 \times 2.5\ m^2$ menggunakan ketiga model tersebut. *Mean Squared Error of Prediction* (MSEP) dihitung dengan rumus:

$$\mathrm{MSEP} = \frac{\sum_{i=1}^{n\_data\ testing} [y_i - \hat{y}_i]^2}{n\_data\ testing}, \dots\dots\dots (3)$$

dimana y_i adalah data aktual (*true value*) dan $\hat{y}_i$ adalah data dugaan**.** Model terbaik dipilih berdasarkan rata-rata MSEP terkecil dari 1000 ulangan.

Setelah diperoleh model terbaik, dilakukan perbandingan rata-rata produktivitas padi antara teknik menghapus data dan teknik imputasi data non-respons. Teknik yang memberikan hasil lebih mirip dengan rata-rata pada data lengkap maka teknik tersebut lebih baik.

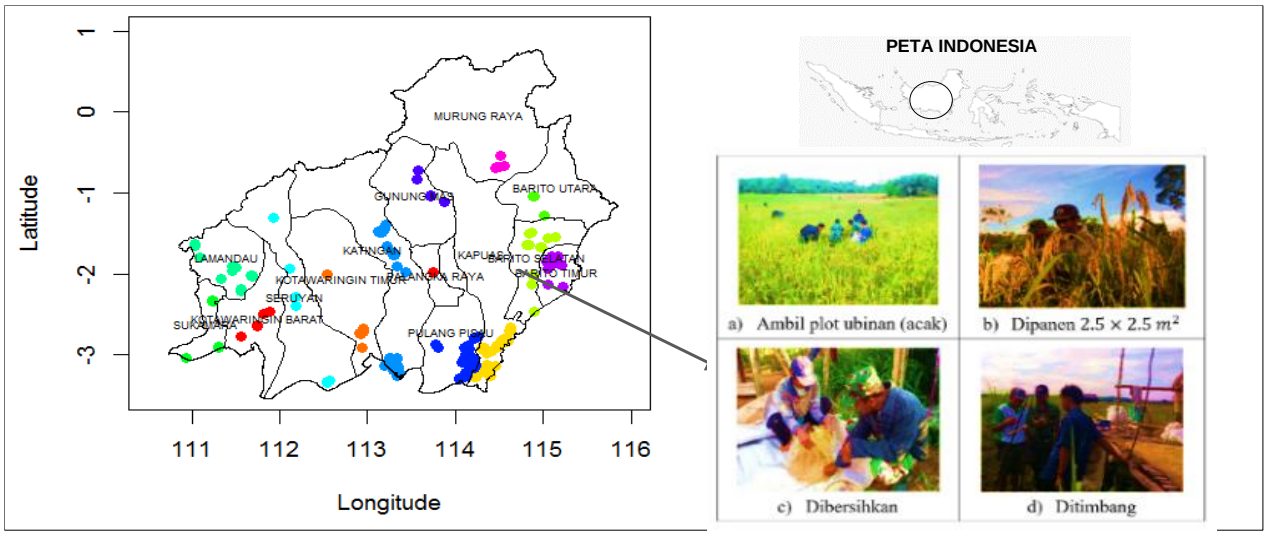
Evaluasi dilakukan dengan membandingkan tingkat presisi metode imputasi melalui tiga skenario data non-respons. Data lengkap dibuat seolah-seolah terdapat data non-respons sebanyak 30 persen. Teknik menghapus data hilang artinya rata-rata produktivitas dihitung dari 70 persen data. Teknik imputasi artinya menduga 30 persen data hilang dengan model yang diajukan kemudian digabung dengan 70 persen data respon. Kemudian digabung dan dihitung rata-rata produktivitas padi. Data 30 persen non-respons diatur dengan skenario tiga kasus, yaitu (1) data hilang terjadi pada kelompok dengan produktivitas padi tinggi, (2) data hilang menyebar secara acak, dan (3) data hilang terjadi pada kelompok dengan produktivitas padi rendah.

Data dan Sumber Data

Data yang digunakan untuk pemodelan adalah data hasil Survei Ubinan di Kalimantan Tengah Tahun 2019. Data dikumpulkan oleh petugas dari Badan Pusat Statistik dibantu penyuluh lapangan dari Dinas Pertanian. Data dikumpulkan saat petani sedang memanen tanaman padi pada petak terpilih. Peubah penelitian dapat dilihat pada Tabel 1.

Tabel 1. Peubah yang digunakan untuk pemodelan

Peubah	Nama Peubah	Sumber berasal dari Survei Ubinan 2019	Keterangan
y	Berat gabah kering panen	Rincian 701	Peubah respon
x_1	Varietas benih (Hibrida atau inbrida)	R609	
x_2	Banyaknya pupuk urea (kg/ha)	R610_1 per R604 kali 10000	
x_3	Banyaknya TSP/SP36 (kg/ha)	R610_2 per R604 kali 10000	
x_4	Banyaknya pupuk KCL (kg/ha)	R610_3 per R604 kali 10000	
x_5	Banyaknya pupuk NPK (kg/ha)	R610_4 per R604 kali 10000	Pengaruh tetap
x_6	Banyaknya pupuk kompos (kg/ha)	R610_5 per R604 kali 10000	
x_7	Serangan OPT (terserang atau tidak)	R804b	
x_8	Dampak perubahan iklim (terdampak atau tidak)	R805b	
x_9	Kabupaten	R102	Pengaruh acak*)
x_{10}	Titik Longitude	R303	Pengaruh aditif geospasial
x_{11}	Titik Latitude	R303	



Gambar 1. Sebaran titik sampel Survei Ubinan di Kalimantan Tengah, Tahun 2019

Data lengkap yang terkumpul pada pelaksanaan Survei Ubinan tanaman padi di Kalimantan Tengah Tahun 2019 adalah sebanyak 585 titik plot yang tersebar di 14 kabupaten/kota se-Kalimantan Tengah. Sebaran titik contoh Survei Ubinan di Kalimantan Tengah tahun 2019 dapat dilihat pada Gambar 1.

HASIL DAN PEMBAHASAN

Beberapa titik sampel Survei Ubinan di Kalimantan Tengah terletak pada lokasi yang berdekatan. Untuk menguji adanya ketergantungan spasial maka digunakan Uji Indeks Moran. Nilai indeks moran diperoleh sebesar $I = 0.4423$ dengan p-value 0.000 sehingga cukup bukti untuk mengatakan terjadi ketergantungan spasial pada data produktivitas padi. Semakin dekat lokasi penanaman padi maka akan semakin mirip nilai produktivitas padi yang dihasilkan. Hal ini menjadi argumen dalam memasukkan fungsi pemulus geospasial ke dalam model. Hasil pendugaan parameter pengaruh tetap dapat dilihat pada Tabel 2.

Tabel 2 menunjukkan bahwa hasil dugaan parameter pengaruh tetap pada ketiga model adalah mirip. Faktor yang berpengaruh nyata terhadap produktivitas padi di Kalimantan Tengah adalah x_1 (Varietas benih), x_2 (Pupuk urea), x_3 (Pupuk TSP/SP36), x_5 (Pupuk NPK), x_7 (Terkena serangan OPT atau tidak), x_8 (terkena dampak perubahan iklim atau tidak). Sedangkan x_4 (Pupuk KCL) dan x_6 (Pupuk kompos) tidak berpengaruh nyata terhadap produktivitas padi di Kalimantan Tengah. Dugaan ragam pengaruh acak dari ketiga model dapat dilihat pada Tabel 3.

Terlihat pada Tabel 3 bahwa saat dilakukan penambahan fungsi pemulus maka dugaan ragam pengaruh acak kabupaten menjadi menyusut. Nilai Dugaan ragam fungsi pemulus $\widehat{\sigma_v^2}$ lebih besar dibanding dugaan ragam pengaruh acak kabupaten $\widehat{\sigma_u^2}$. Artinya, penambahan fungsi pemulus dapat meningkatkan kinerja prediksi model. Perbandingan nilai MSEP antarmodel disajikan pada Gambar 2.

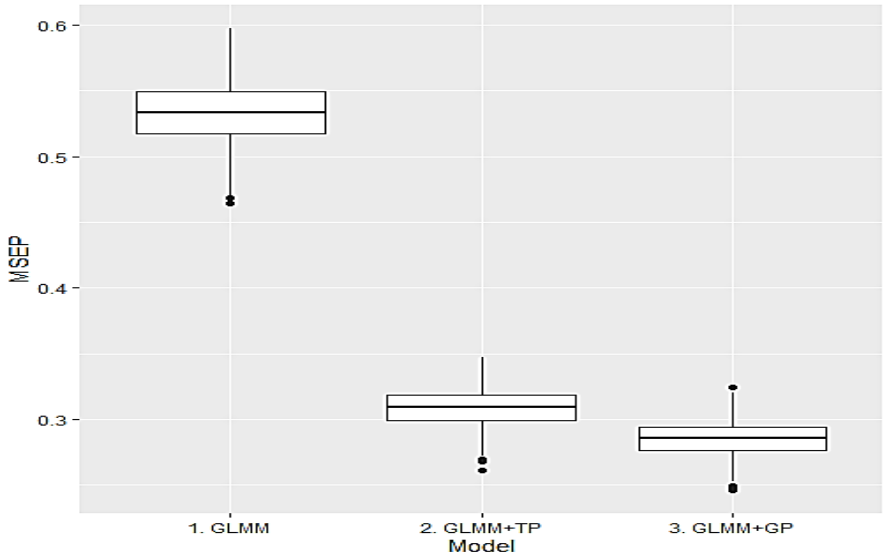
Dapat dilihat dari Gambar 2 bahwa GLMM+GP memiliki nilai MSEP terkecil disusul GLMM+TP dan GLMM. Nilai Quartil 1 (Q1) pada MSEP GLMM+TP lebih besar dibanding nilai Q3 pada MSEP GLMM+GP. Hal ini mengindikasikan bahwa model GLMM+GP memiliki performa terbaik secara nyata. Dengan pengulangan 1000 kali maka derajat bebas galat sebesar $1000 - 1 = 999$ yang menyebabkan kuadrat tengah galat menjadi sangat kecil mendekati nol. Oleh karena itu, dapat ditunjukkan bahwa model GLMM+GP merupakan model yang

Tabel 2. Nilai dugaan parameter pada model GLMM, GLMM+TP, dan GLMM+GP

Peubah	Dugaan parameter			Pr(>|t|)		
	GLMM	GLMM+TP	GLMM+GP	GLMM	GLMM+TP	GLMM+GP
Intersep	2.2800	2.4470	2.4402	0.0000	0.0000	0.0000
x_1	-0.7901	-0.5452	-0.5144	0.0000	0.0000	0.0000
x_2	0.0015	0.0016	0.0014	0.0034	0.0004	0.0016
x_3	0.0027	0.0022	0.0020	0.0000	0.0006	0.0023
x_4	-0.0010	-0.0006	-0.0011	0.6850	0.7680	0.6014
x_5	0.0024	0.0014	0.0013	0.0000	0.0000	0.0002
x_6	0.0005	0.0007	0.0008	0.2808	0.0980	0.0796
x_7	0.2680	0.1331	0.1413	0.0000	0.0246	0.0175
x_8	0.1574	0.2076	0.2004	0.0259	0.0011	0.0018

Tabel 3. Dugaan ragam pengaruh acak pada GLMM, GLMM+TP, GLMM+GP

Pengaruh Acak	GLMM	GLMM+TP	GLMM+GP
$\widehat{\sigma_u^2}$	0.2017	0.0006	0.0064
$\widehat{\sigma_v^2}$	-	10.5924	2.6835
$\widehat{\sigma_e^2}$	0.5515	0.5904	0.5740



Gambar 2. Perbandingan MSEP antara GLMM, GLMM+TP, dan GLMM+GP

memberikan performa terbaik secara nyata dalam pendugaan produktivitas padi dibanding dua model lainnya. Model GLMM+GP dapat digunakan untuk mengimputasi data hilang berat gabah pada Survei Ubinan pada pelaksanaan Survei Ubinan berikutnya.

Kajian data non-respons

Pada bagian ini digunakan data asli hasil Survei Ubinan 2019 di tiga kabupaten penghasil padi terbesar di Kalimantan Tengah (Kapuas, Pulang Pisau, dan katingan). Langkah yang diterapkan pada kajian ini adalah sebagai berikut: Pertama, data ubinan yang lengkap dibagi menjadi 75 persen data respon dan 25 persen data non-respon. Pengambilan 25 persen data non-respon dilakukan secara acak dengan tiga skenario, yaitu data hilang terjadi pada kelompok dengan produktivitas padi tinggi (kasus I), data hilang menyebar secara acak (kasus II), dan data hilang terjadi pada kelompok dengan produktivitas padi yang rendah (kasus III). Langkah kedua, pemodelan dilakukan menggunakan data lengkap. Ketiga, menduga produktivitas padi menggunakan model pada tahap dua. Selanjutnya, mengulangi ketiga langkah tersebut sebanyak 200 iterasi.

Kajian ini menunjukkan bahwa teknik imputasi lebih baik dibanding teknik menghapus data non-respons. Teknik mengabaikan non-respon (*listwise deletion*)

Tabel 4. Perbandingan rata-rata berat gabah seluas $2.5 \times 2.5\ m^2$ ($\bar{y}$) antara teknik mengabaikan data non-respons, teknik imputasi dengan GLMM+GP, dan data lengkap

Jenis data hilang	$\bar{y}$ pada data lengkap	$\bar{y}$ dengan mengabaikan non-respons		$\bar{y}$ setelah ditambah hasil imputasi	
		dugaan	selisih *(2) - (3)*	dugaan	selisih *(2) - (5)*
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Kasus I[1)]					
Katingan	1.96	1.49	0.47	1.61	0.35
Pulang Pisau	2.23	1.74	0.49	1.85	0.38
Kapuas	2.58	2.15	0.43	2.24	0.34
Kasus II[2)]					
Katingan	1.96	2.01	-0.05	1.94	0.02
Pulang Pisau	2.23	2.19	0.04	2.21	0.02
Kapuas	2.58	2.54	0.04	2.54	0.04
Kasus III[3)]					
Katingan	1.96	2.39	-0.43	2.22	-0.26
Pulang Pisau	2.23	2.68	-0.45	2.54	-0.31
Kapuas	2.58	2.95	-0.37	2.79	-0.21

hanya bisa diterapkan dalam kasus data hilang tersebar secara acak atau *Missing at Random* (MAR). Dalam praktiknya, pada pelaksanaan Survei Ubinan di Kalimantan Tengah, seringkali non-respons terjadi tidak secara acak. Oleh karena itu, metode imputasi disarankan untuk diterapkan dalam mengatasi missing data yang disebabkan non-respons.

Dapat dilihat pada Tabel 4 bahwa saat data hilang tersebar secara acak (kasus II) maka angka rata-rata berat gabah mirip dengan rata-rata saat tidak ada data hilang. Artinya, saat data hilang tersebar acak maka metode mengabaikannya dapat diterapkan. Berbeda dengan jenis data hilang pada kasus I dan III, rata-rata produktivitas yang dihasilkan bisa *underestimate* ataupun *overestimate* sehingga perlu ditangani. Terlihat dalam kasus I dan III atau kasus MNAR (*Missing Not at Random*), teknik imputasi data hilang menghasilkan rata-rata yang lebih mirip dengan rata-rata data lengkap. Dapat ditunjukkan bahwa metode imputasi bermanfaat saat kondisi missing values adalah MNAR karena mampu mempersempit bias dugaan.

KESIMPULAN

Terdapat tiga faktor utama yang berpengaruh nyata terhadap produktivitas padi di Kalimantan Tengah tahun 2019 yaitu varietas, pupuk, dan faktor lingkungan. Pupuk yang berpengaruh nyata adalah pupuk urea, TSP/SP36, dan NPK/ pupuk majemuk, sedangkan faktor lingkungan yang berpengaruh nyata adalah serangan OPT dan dampak perubahan iklim. Berdasarkan indeks Moran, terjadi ketergantungan spasial pada data produktivitas padi di Kalimantan Tengah. Semakin dekat lokasi penanaman padi maka akan semakin mirip nilai produktivitas padinya. Hal ini menjadi argumen dalam memasukkan fungsi pemulus geospasial ke dalam model. Hasil perbandingan kinerja antarmodel menunjukkan bahwa model dengan penambahan fungsi pemulus *Gaussian Process* memberikan kinerja terbaik dengan nilai MSEP tekecil secara signifikan. Berdasarkan kajian, dapat ditunjukkan bahwa metode imputasi bermanfaat saat data non-respons menyebar tidak acak atau kondisi missing valuesnya adalah MNAR. Sedangkan pada kasus MAR teknik *listwise deletion* sama baiknya dibanding teknik imputasi. Bardasarkan hasil penelitian ini, saat gagal melakukan pengukuran berat gabah, petugas BPS disarankan tetap melakukan wawancara kepada petani untuk memperoleh peubah-peubah yang dapat dimanfaatkan untuk mengimputasi data hilang.

DAFTAR PUSTAKA

Ardiansyah, M., & Tofri, Y. (2019). Perbandingan Data Produktivitas Padi Antara Hasil Wawancara Pascapanen dengan Data Survei Ubinan di Kalimantan Tengah. *Jurnal Penelitian Pertanian Tanaman Pangan, 3*(1), 17-22. doi:http//dx.doi.org/10.21082/jpptp.v3n1.

Ardiansyah, M., Buana, W. P., & Kurnia, A. (2020). Prediksi Produktivitas Padi Melalui Survei Ubinan Menggunakan Model Linier dan Quantile Regression Forest. *Jurnal Penelitian Pertanian Tanaman Pangan, 4*(3), 135-144. doi:http//dx.doi.org/10.21082/jpptp.v4n3.

Ardiansyah, M., Kurnia, A., Sadik, K., Djuraidah, A., & Wijayanto, H. (2021). Numerical Prediction of paddy weight of Crop Cutting Survey using Generalized Geoadditive Linear Mixed Model. *Journal of Physics: Conference Series, 1863 (2021) 012024*, 1-17. doi:10.1088/1742-6596/1863/1/012024

Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association, 88*(421), 9–25.

Chhabra, G., Vashisht, V., & Ranjan, J. (2019). A Review on Missing Data Value Estimation Using Imputation Algorithm. *Journal of Advanced Research in Dynamical and Control Systems, 11*(07), 312-318.

Curley, C., Krause, R. M., Feiock, R., & Hawkins, C. V. (2019). Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database.

Urban Affairs Review, 55(2), 591-615. doi:10.1177/1078087417726394.

Djuraidah, A. (2020). *Monograf Penerapan dan Pengembangan Regresi Spasial dengan Studi Kasus pada Kesehatan, Sosial, dan Ekonomi.* Bogor: IPB Press.

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. Statistical Science. *1*(3), 297-318.

Lin, X., & Zhang, D. (1999). 1999. Inference in generalized additive mixed models by using smoothing splines. *Royal Statistical Society, 61*(2), 381-400.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, 135*(3), 370-384.

Vo, G., & Pati, D. (2017). Sparse Additive Gaussian Process with Soft Interactions. *Open Journal of Statistics, 7*, 567-588. doi:10.4236/ojs.2017.74039.

Wood, S. N. (2017). *Generalized Additive Models An Introduction with R Second Edition.* London: Chapman & Hall/CRC Press, Taylor & Francis Group.

Wood, S. N., Scheipl, F., & Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Stat Comput, 23*, 341-360. doi:10.1007/s11222-012-9314-z.

Yu, S., Wang, G., Wang, L., Liu, C., & Yang, L. (2019). Estimation and Inference for Generalized Geoadditive Models. *Journal of the American Statistical Association*, 1-14. doi:10.1080/01621459.2019.1574584.

.