BAB II

TINJAUAN PUSTAKA

2.1 Definisi Pengangguran

Pengangguran merupakan hal yang akan selalu muncul didalam perekonomian, dimana saat pengeluaran agregatnya lebih rendah dibandingkan dengan kemampuan faktor-faktor produksi yang telah tersedia didalam perekonomian untuk dapat menghasilkan barang-barang dan juga jasa (Prasaja, 2013). Navarrete menjelaskan dalam bukunya "*Underemployment in Underdeveloped Countries*" pengangguran dapat dilukiskan sebagai suatu keadaan dimana adanya pengalihan sejumlah faktor tenaga kerja ke bidang lain yang mana tidak akan mengurangi output keseluruhan sektor asalnya atau dikatakan bahwa peoduktivitas marginal unit-unit faktor tenaga tempat asal mereka bekerja adalah nol atau hampir mendekati nol atau juga negatif (Jhingan, 2014:22).

Salah satu alasan pengangguran selalu muncul didalam pengangguran adalah pencarian kerja. Pencarian kerja (*job search*) adalah suatu proses seseorang untuk mencocokkan pekerja dengan pekerjaan yang sesuai dengan bakat dan juga keterampilan sesuai yang dimiliki oleh mereka. Namun, jika semua pekerja dan pekerjaan tidak ada bedanya, maka tidak menutup kemungkinan bagi para pekerja bahwa mereka cocok dengan pekerjaan apa saja, akan tetapi pada kenyataannya bakat dan juga kemampuan seseorang itu berbeda-beda (Mankiw dkk, 2012).

Definisi pengangguran adalah angkatan kerja yang tidak memiliki pekerjaan, dan pengangguran terbuka adalah pengangguran sukarela, atau sengaja menganggur untuk mendapatkan pekerjaan yang lebih baik. Seseorang baru dikatakan menganggur bila dia ingin bekerja dan telah berusaha mencari kerja, namun tidak mendapatkannnya. Dalam ilmu kependudukan (demografi), orang yang mencari kerja masuk dalam kelompok penduduk yang disebut angkatan kerja. Berdasarkan kategori usia, usia angkatan kerja adalah 15-64 tahun, tetapi tidak semua orang yang berusia 15-64 tahun dihitung sebagai angkatan kerja (Zurisdah, Z 2016).

Definisi pengangguran menurut BPS pengangguran terbuka (*open unemployment*) didasarkan pada konsep seluruh angkatan kerja yang mencari perkerjaan, baik yang mencari perkerjaan pertama kali maupun yang pernah bekerja sebelumnya. Sedang pekerja yang digolongkan setengah penganggur *(underemployment)* adalah pekerja yang masih mencari pekerjaan penuh atau sambilan dan mereka yang bekerja dengan jam kerja rendah. Setengah pengaggur sukarela adalah setengah penganggur tetapi tidak mencari pekerjaan atau tidak bersedia menerima pekerjaan lain. Setengah penganggur terpaksa adalah setengah penganggur yang masih mencari pekerjaan atau bersedia menerima pekerjaan. Pekerja digolongkan setengah penganggur parah (*severe underemployment)* apabila ia masuk setengah menganggur dengan jam kerja kurang dari 25 jam seminggu.

a. Pengangguran dalam Sektor Informal

Pengangguran terbuka biasanya terjadi pada generasi muda yang baru menyelesaikan pendidikan menengah dan tinggi. Ada kecenderungan mareka yang baru menyelesaikan pendidikan berusaha mencari kerja sesuai dengan aspirasi mareka. Aspirasi mareka biasanya adalah bekerja di sektor modern atau di kantor, untuk mendapatkan pekerjaan itu mareka bersedia menunggu untuk beberapa lama, tidak tertutup kemungkinan mereka berusaha mencari perkerjaan itu di kota atau di provinsi atau daerah yang kegiatan industri telah berkembang. Hal ini menyebabkan angka pengangguran tinggi di perkotaan atau di daerah kegiatan industri atau sektor modern berkembang. Sebaliknya pengangguran terbuka rendah di daerah atau provinsi yang tumpu pada sektor pertanian. Hal tersebut penyediaan pekerjaan di sektor informal oleh sebab rendahnya pendidikan dan kurang menjamin kelangsungan hidup.

b. Pengukuran Tingkat Pengangguran

Badan statistik negara mengelompokkan orang dewasa pada setiap rumah tangga yang disurvei ke dalam satu kategori berikut.

1. Bekerja
2. Pengangguran
3. Tidak termasuk angkatan kerja

Setelah mengelompokkan seluruh individu yang disurvei ke dalam tiga kategori tersebut, badan statistik negara menghitung berbagai statistik untuk merangkum kondisi angkatan kerja. Angkatan kerja *(labor force*) adalah jumlah orang yang berkerja dan tidak berkerja.

Angkatan kerja = Jumlah orang yang bekerja + Jumlah yang tidak bekerja.

Tingkat pengangguran *(unemployment rate)* adalah persentase angkatan kerja yang tidak bekerja:

$$\text{Tingkat Pengangguran} = \frac{jumla\ h\ pengangguran}{jumla\ h\ angkatan\ \ kerja}\ x\ 100 \qquad (2.1)$$

Setelah itu, tingkat pengangguran untuk seluruh populasi penduduk dewasa dan untuk kelompok yang lebih sempit, seperti laki-laki dan perempuan dapat dihitung.

2.2 Ketenagakerjaan

2.2.1 Penduduk Usia Kerja

Penduduk usia kerja didefenisikan sebagai penduduk yang berumur 15 tahun dan lebih. Mereka terdiri dari "Angkatan Kerja" dan "Bukan Angkatan Kerja". Proporsi penduduk tergolong "Angkatan Kerja" adalah mereka yang aktif dalam kegiatan ekonomi. Keterlibatan penduduk dalam kegiatan ekonomi diukur dengan porsi penduduk yang masuk dalam pasar pekerjaan. Tingkat Partisipasi Angkatan Kerja (TPAK) merupakan ukuran yang menggambarkan jumlah angkatan kerja untuk setiap 100 penduduk usia kerja.

Penduduk Jawa Barat berusia 15 tahun atau lebih pada tahun 2017 mencapai 35,35 juta orang. jumlah angkatan kerja sebanyak 22,39 juta orang, dimana 20,53 juta orang diantaranya bekerja di berbagai sektor usaha, sedangkan sisanya 1,84 juta masih menganggur. Julah tersebut menjadikan angka tingkat pengangguran terbuka menjadi 8,22%. Penduduk usia produktif (15-64 tahun)

mencapai 32,67 juta orang, dan usia nonproduktif sebanyak 15,36 juta menjadikan angka *dependency ratio* atau rasio ketergantungan menjadi 47,02 yang artinya dalam 100 orang usia produktif menanggung 47 orang usia nonproduktif. Nilai menunjukkan bahwa Jawa Barat telah memasuki periode bonus demografi dimana 1 orang usia nonproduktif ditanggung oleh setidaknya 2 orang usia produktif.

2.2.2 Komposisi Penduduk yang Bekerja

Perekonomian Jawa Barat diperkirakan digerakkan oleh setidaknya 20,55 juta orang bekerja. Mereka berkerja di berbagai lapangan usaha yang ada. Sebagian besar atau 28,64 persen, dan sektor jasa 10,91 persen. Pekerja di Jawa Barat didominasi oleh lulusan SD, yakni mencapai 30,17 pesen, dan pekerja lulusan SMA ke atas mencapai 40,87 persen.

2.3 Pendugaan Area Kecil (*Small Area Estimation*)

SAE adalah salah satu teknik statistik yang digunakan untuk menduga parameter subpopulasi dengan ukuran sampel yang relatif kecil. Teknik ini mengembangkan data survei dan sensus untuk mengestimasi tingkat kesejahteraan atau indikator lainnya untuk unit geografis seperti kecamatan atau pedesaan (Davies, 2003). Suatu daerah disebut *small area* jika daerah tersebut jumlah contoh yang terambil kurang besar untuk mendapatkan nilai pendugaan langsung yang akurat. Nilai pendugaan langsung pada area kecil merupakan penduga tak bias tetapi memiliki ragam yang besar karena diperoleh dari ukuran contoh yang kecil (Ramsini *et.al* (2001) dalam Kurnia dan Notodiputro (2006)).

SAE merupakan pendugaan suatu area yang lebih kecil dengan memanfaatkan informasi dari luar area. Informasi dari dalam area itu sendiri, dan

dari luar survey (Longford. 2005). Terdapat dua masalah pokok dalam pendugaan area kecil. Masalah pertama adalah bagaimana menghasilkan suatu dugaan parameter yang cukup baik dengan ukuran sampel yang kecil pada suatu area kecil. Masalah kedua adalah bagaimana menduga *Mean Square Error* (MSE). Solusi untuk masalah tersebut adalah dengan "meminjam informasi" dari dalam area, luar daerah, maupun luar survei (Pfefferman 2002).

Pendugaan parameter pada suatu area kecil dapat dilakukan dengan pendugaan secara langsung (*direct estimatoin*) maupun pendugaan secara tidak langsung (*indirect estimation*). Hasil pendugaan langsung pada suatu daerah kecil merupakan penduga tak bias meskipun memiliki varian yang besar dikarenakan dugaannya diperoleh dari ukuran sampel yang kecil (Ramsini *et al*. 2001). Sedangkan tak langsung merupakan pendugaan dengan cara memanfaatkan informasi dari variabel lain yang berhubungan dengan parameter yang diamati.

Proses pendugaan pada suatu area subpopulasi dapat dibagi menjadi dua macam, yaitu:

1. Penduga Berbasis Rancangan

Rao (2003) menyebutkan bahwa pendugaan pada metode berbasis rancangan merupakan pendugaan pada suatu area berdasarkan data contoh dari area tersebut. pada proses pendugaan tersebut dapat digunakan informasi tambahan (*auxiliary informaton*) untuk menduga parameter yang menjadi perhatian. Pendekatan yang digunakan pada proses pendugaan ini adalah pendekatan berbasis rancangan. Pada pendugan ini diasumsikan terjadi galat pengukuran.

2. Penduga Berbasis Model

Pendugaan pada metode berbasis model merupakan pendugaan pada suatu area dengan cara menghubungkan informasi pada area dengan area lain melalui model yang tepat. Hal ini berarti bahwa dugaan tersebut mencakup data dari area lain (Kurnia & Notodiputro 2006). Pendugaan tidak langsung (*indirect estimation*) dilakukan dengan cara memanfaatkan informasi peubah lain yang berhubungan dengan parameter yang diamati. Contoh informasi yang dapat digunakan adalah catatan sensus ataupun survei pada area tersebut. Pendugaan tidak langsung baerdasarkan model area kecil (*small area model*) dikatakan sebagai penduga berbasis model (Rao 2003). Ramsini *et al.*(2001) menyatakan bahwa penduga tidak langsung yang diperoleh dengan memanfaatkan informasi peubah lain yang berhubungan dengan parameter yang diamati sering disebut sebagai penduga berbasis model adalah metode EB (*Empirical Bayes*), EBLUP (*Empirical Best Linear Unbiased*), dan HB (*Hierarchical Bayes*).

2.3.1 Model Area kecil

Terdapat dua ide utama yang digunakan untuk mengembangkan model pendugaan parameter area kecil yaitu:

1. Model pengaruh tetap (*fixed effect model*) dimana asumsi bahawa keragaman di dalam area kecil, variabel respon dapat diterangkan seluruhnya oleh hubungan keragaman yang bersesuaian pada informasi tambahan.
2. Pengaruh acak area kecil (*random effect*) dimana asumsi keragaman spesifik area kecil tidak dapat ditearangkan oleh informasi tambahan.

Gabungan daru kedua asumsi tersebut membentuk suatu model pengaruh campuran (*mixed model*). Oleh karena variabel respon diasumsika berdistribusi normal maka penduga area kecil yang dikembangkan merupakan bentuk khusus dari *General Linear Mixed Model* (GLMM).

Model *small area* biasanya menggunakan model linear campuran dalam betuk

$$y = Xb + Zu + e \tag{2.2}$$

dimana *X* adalah matrix peubah penyerta, *Z* adalah vektor acak yang biasa dikenal sebagai pengaruh area kecil, dan *e* adalah vektor dari galat sampel (Rao, 2003). Menurut Rao (2003) ada dua model dasar pendugaan area kecil, yaitu *basic area level model* dan *basic unit level model*.

1. Model berbasis area level

Model berbasis area level merupakan model yang didasarkan pada ketersediaan data pendukung yang hanya ada untuk level area tertentu, misalkan $x_i = \left(x_{li},.....,x_{pi}\right)^T$ dengan parameter yang akan diduga adalah θ_i yang diasumsikan mempunyai hubungan dengan x_i (Rao, 2003). Data pendukung tersebut digunakan untuk membangun model θ_i adalah:

$$\theta_i = x_i^T \beta + v_i, i = 1,...., \mathrm{m} \tag{2.3}$$

Dimana m adalah bayaknya area dengan $\beta = (\beta_1,....,\beta_p)^T$ merupakan vektor px1 koefisien regresi untuk variabel penyerta x_i dan v_i adalah pengaruh acak area kecil yang diasumsikan berdistribusi N(0, σ_v^2).

Dapat diketahui estimator θ_i dengan mengasumsi bahwa model penduga langsung $\hat{\theta}_i$ telah tersedia, yaitu:

$$\hat{\theta}_i = \theta_i + e_i, i = 1,....,\text{m} \quad (2.4)$$

dengan $e_i = \text{N}(0, \psi i)$ dan ψi diketahui.

Gabungan antara dua model (2.1) dan (2.2) akan mengahsilkan persamaan model gabungan (*mixed model*) yang dikenal dengan model Fay-Herriot (Fay dan Herriot, 1979).

$$\hat{\theta}_i = x_i^T \beta + v_i + e_i, i = 1,...,\text{m} \quad (2.5)$$

Dimana x_i adalah vektor pxl variabel penyerta tingkat area $v_i \sim \text{N}(0, \sigma_i^T)$ dan $e_i \sim \text{N}(0, \psi i)$, dengan varian ψi, yang diketahui dari data dimana v_i dan e_i saling bebas.

Dimana keragaman variabel respon di dalam area kecil di asumsikan dapat diterangkan oleh hubungan variabel respon dengan informasi tambahan (variabel prediktor) yang disebut dengan model pengaruh tetap (*fixed effek models*). Selain terdapat komponen keragaman spesifik area kecil yang tidak bisa diterangkan oleh informasi tambahan (variabel prediktor), disebut dengan komponen pengaruh acak area kecil (*random effect*). Gabungan dua asumsi tersebut membentuk model pengaruh acak campuran atau model linear campuran (Kurnia, 2009).

Saei dan Chambers (2003) mengemukakan dua ide utama dalam mengembangkan model SAE yaitu (1) asumsi bahwa keragaman didalam subpopulasi peubah respon dapat diterangkan seluruhnya oleh hubungan keragaman yang bersesuaian pada informasi tambahan, disebur model pengaruh

tetap (*fixed effect*), (2) asumsi keragaman spesifik subpopulasi tidak dapat diterangkan oleh informasi tambahan dan merupakan pengaruh acak subpopulasi (*random effect*). Gabungan daru kedua asumsi tersebut membentuk suatu pengaruh campuran (*mixed effect*). Terjadi kelemahan jika model yang dibuat tidak menggambarkan kondisi wilayah/daerah yang sebenarnya.

2. Model berbasis unit level

Model berbasis unit level merupakan suatu model dimana data-data pendukung yang tersedia bersesuaian secara individu dengan data respon, nilai $x_i = (x_{ij1},x_{ij2},...,x_{ijp})^T$, sehingga didapat suatu model regresi tesarang:

$$y_{ij} = x_{ij}^T \beta + v_i + e_{ij} \text{ , } i=1,....,m \text{ dan } j=1,..., n_i \qquad (2.6)$$

Dimana j adalah banyaknya gizi buruk pada daerah ke-i dengan $v_i \sim N(0, \sigma_v^2)$ dan $e_i \sim N(0, \sigma_v^2)$.

Dimana $x_{ij} = (x_{1j1},...,x_{ijp})^T$ yang merupakan data penyerta unit tertentu, *p* adalah variabel prediktor, *j* adalah angka pengangguran pada area ke-*i*, dan v_i = pengaruh acak area yang diasumsikan merupakan variabel yang bersifat *iid*

$$e_{ia} = k_{ia} \times \tilde{e}_{ia} \qquad (2.7)$$

Dimana

k_{ia} : Konstanta

$\tilde{e}_{ia}$: variabel acak yang bersifat *iid* dan bebas terhadap v_i, dimana $E_m(\tilde{e}_{ia}) = 0$ dan

$$V_{\varepsilon m}(\tilde{e}_{ia}) = \sigma_\varepsilon^2$$

v_i dan e_{ia} seringkali diasumsikan memiliki distribusi peluang normal

Perbedaan mendasar pada kedua model tersebut yaitu pada penggunaan data pendukung yang tersedia. Pada model SAE level area, data pendukung yang tersedia hanya untuk level area tertentu. Model ini menghubungkan estimator langsung dengan variabel penyerta dari domain lain untuk setiap area. Sedangkan model level unit mengasumsikan bahwa variabel penyerta yang tersedia bersesuaian secara individu dengan variabel respon.

Penelitian ini menggunakan model *basic area level model* karena data pendukungnya hanya ada untuk level area tertentu, yaitu pada level kabupaten/kota. Model berbasis area dengan satu peubah penyerta, model (2.2) bisa dinyatakan sebagai:

$$y_i = \theta_i + e_i \tag{2.7}$$

$$\hat{\theta}_i = x_i^T \beta + b_i u_i + e_i \tag{2.8}$$

Dengan β merupakan vektor koefisien regresi untuk data pendukung x_i dengan u_i berdistribusi independen N(0, σ_v^2), sebagai pengaruh acak yang diasumsikan normal dan e_i ~ N(0, ψi) (Fay dan Herriot, 1997)

2.4 Model Regresi *Spline*

Regresi nonparametrik merupakan suatu metode statistika yang digunakan untuk mengetahui hubungan antara variabel respon dan prediktor yang tidak diketahui bentuk fungsinya, hanya diasumsikan fungsi *smooth* (mulus) dalam arti termuat dalam suatu ruang fungsi tertentu, sehingga regresi nonparametrik memiliki fleksibilitas yang tinggi (Eubank, 1988). Model regresi nonparametrik secara umum dapat disajikan sebagai berikut:

$$y_i = m(x_i) + e_i, 1,2,...,n \quad (2.9)$$

Dengan y_i adalah variabel respon, fungsi $m(x_i)$ adalah dungsi yang *smooth* dimana tidak diketahui bentuknya. Variabel x_i sebagai variabel prediktor dengan $e_i \sim N(0,\sigma^2)$. Agar dapat menangani struktur $m(x_i)$ yang tidak linear, didefenisikan K buah knot $k_1,...,k_k$ dan dengan mengambil basis fungsi *polynomial* terputus diperoleh model berikut:

$$m(x_i) = \beta_0 + \beta_i x_i + ... + \beta_p x_i^p + \sum_{j=i}^{k} \gamma_i (x_i - k_i)_+^p,$$

Dengan p adalah derajat spline, $(x_i - k_j)_+ = maks\{0,(x_i - k_i)\}, k_j$ dimana $j = 1,...,K$ merupakan himpunan titik knot. Dengan menetapkan $\boldsymbol{\beta} = (\beta_0,...,\beta_p)^T$ adalah vektor koefisien parametrik. $\boldsymbol{\gamma} = (\gamma_1,...,\gamma_k)^T$ adalah vektor koefisien spline, = [1 x_i ... x_i^p x_i]$_{1\le i\le n}$, $Z = \left[(x_i - k_1).....(x_i = k_K)_+^p\right]_{1\le i\le n}$, dengan

$$(x_i - k_j)_+^p \begin{cases} = (x_i - k_j)_+^p & untuk\ x_i \ge k_j \\ = 0 & untuk\ x_i < k_j \end{cases}$$

Sehingga model (2.5) dapat ditulis sebagai berikut:

$$y_i = \beta_0 + \beta_i x_i + \cdots + \beta_p x_i^p + \sum_j^k \gamma_j (x_i - k_j)_+^p + e_i$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{e}$$

dimana $\mathbf{Y} = (y_i,... , y_k)^T$ (2.10)

Model (2.10) disebut sebagai *regresi spline smoothing* (Opsomer, 2004) dari bentuk matematis fungsi *spline* pada model tersebut menunjukkan bahwa *spline* merupakan model optimal terputus, tetapi masih bersifat kontinu pada knot-knotnya.

Knot dapat diartikan sebagai suatu titik fokus dalam fungsi *spline* sedemikian sehingga kurva yang dibentuk tersegmen pada titik tersebut. Titik knot merupakan salah satu hal yang sangat penting dalam pendekatan *spline*. Strategi yang digunakan untuk memilih dan menentukkan lokasi knot dengan tepat sangat dibutuhkan agar diperoleh model *spline* yang optimal. Jika jumlah knot terlampau banyak maka model yang dihasilkan akan *overfitting*.

Salah satu metode pemilihan titik knot optimal adalah dengan menggunakan *Generalized Cross Validation* (GCV).

Definisi GCV dapat ditulis sebagai berikut:

$$\text{GCV}((K) = \frac{MSE(K)}{\left[n^{-1}trace(1-A(K))\right]^2} \tag{2.11}$$

Dimana $MSE\ (\boldsymbol{K}) = n^{-1}y^t(\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{K}))\,y$, $\boldsymbol{K} = (K_1, K_2, \ldots, K_N)$ adalah titik knot dan matriks $\boldsymbol{A}(\boldsymbol{K})$ diperoleh dari persamaan $\hat{y} = A(\boldsymbol{K})y$.

2.5 Regresi *Penalized Spline*

Regresi *penalized spline* yaitu regresi yang diperoleh berdasarkan kuadrat terkecil (*least square)* dengan penalty kekasaran. *Penalized spline* mempunyai banyak kesamaan dengan smoothing spline, tetapi jenis *penalty* yang digunakan pada *penalized spline* lebih umum dibandingkan pada *smoothing spline* (Ruppert, 2003).

Menurut Hall dan Opsomer (2005), regresi *penalized spline* merupakan suatu pendekatan *smoothing* yang popular karena kesederhanaannya dan fleksibilitasnya. Pemodelan *penalized spline* memberikan pemilihan knot yang fleksibel. Salah satu alternatif untuk mengoptimalkan fit model terhadap data

adalah dengan menambahkan *penalty* pada parameter *spline*. Dengan cara ini dapat dipilih jumlah knot yang banyak dan mencegah *overfitting* dengan menempatkan kendala *(constraint*).

Terdapat dua komponen penting dalam mengestimasi *penalized spline*, yang pertama adalah pemilihan karakter *smoothing*, sementara yang kedua adalah pemilihan jumlah knot dan lokasinya (Yao dan Lee, 2008). Pada persamaan (2.7) dapat dinyatakan ke dalam bentuk matriks yaitu

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{e}$$

Dimana

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_2 \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_k^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_1 & x_n^p \end{bmatrix}, \boldsymbol{Z} = \begin{bmatrix} (x_1 - k_1)_+^p & \cdots & (x_1 - k_K)_+^p \\ \vdots & \ddots & \vdots \\ (x_n - k_1)_+^p & \cdots & (x_1 - k_K)_+^p \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_k \end{bmatrix}, \text{ dan } \boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

Estimator *penalized spline* diperoleh dengan meminimumkan fungsi *Penalized Least Square* (PLS) sebagai berikut:

$$L = \|\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\gamma\|^2 + \lambda \boldsymbol{\gamma}^T \boldsymbol{\gamma} \qquad (2.12)$$

Dengan memisalkan **C** = [***X*** , ***Z***] dan $\ddot{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}$sehingga persamaan (2.12) dapat dituliskan sebagai berikut:

$$L = \left\|\boldsymbol{y} - \boldsymbol{C}\ddot{\theta}\right\|^2 + \lambda \ddot{\theta}^T \boldsymbol{D} \ddot{\theta} \qquad (2.13)$$

Dimana diketahui **D** merupakan matrik *penalty*

$$\boldsymbol{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \vdots & 0 \end{bmatrix} = \begin{bmatrix} 0_{(P+1)\times 2} & 0_{(P+1)\times K} \\ 0_{K\times(P+1} & I_{K\times K} \end{bmatrix}$$

Dengan parameter λ parameter *smoothing*, dimana $\lambda \geq 0$. Suku pertama pada persamaan (2.13) adalah jumlah kuadrat error dan suku keduanya adalah penalty kekasaran. Menurut Djuraidah, et al (2006) Estimator *penalized spline* yang diperoleh adalah

$$\hat{\hat{\theta}} = (\boldsymbol{C^T C} + \boldsymbol{\lambda D})^{-1}\, \boldsymbol{C^T y} \tag{2.14}$$

dengan demikian didapatkan.

$$\hat{\boldsymbol{y}} = \boldsymbol{C}(\boldsymbol{C^T C} + \boldsymbol{\lambda D})^{-1}\, \boldsymbol{C^T y} \tag{2.15}$$

Berdasarkan uraian di atas, nilai $\hat{\theta}$ bergantung pada parameter *smoothing* λ. Jika nilai λ besar akan menghasilkan bentuk kurva regresi yang sangat halus. Sebaliknya, jika nilai λ kecil akan memberikan bentuk kurva regresi yang sangat kasar. Akibatnya pemilihan parameter *smoothing* optimal perlu dilakukan. Dengan menggunakan *Generalized Cross-Validation* (GCV) yang didefinisikan sebagai berikut:

$$GCV(\lambda) = \frac{n^{-1} RSS(\lambda)}{\left(1 - n^{-1} df\lambda\right)^2} = \frac{MSE(\ )}{(n^{-1} tr(I - S_\lambda))^2} \tag{2.16}$$

Dimana $RSS(\lambda) = \sum_{i=1}^{n} \left(y_1 - \hat{y}_i\right)^2, df_\lambda = tr(\boldsymbol{S}_\lambda)$

$\boldsymbol{S}_\lambda = \boldsymbol{C}(\boldsymbol{C}^T \boldsymbol{C} + \boldsymbol{\lambda D})^{-1}\, \boldsymbol{C^T}$ yang disebut dengan matriks *smooting* (Ruppert, et al., 2003; Griggs, 2013)

Pada penelitian ini untuk melakukan penetuan jumlah titik knot dapat dilakukan dengan metode *fixed selection method*. Tujuan utama untuk semua metode pemilihan knot K adalah untuk memastikan bahwa K cukup besar agar lebih fleksibel ketika mengontrol kemulusan kurva yang diestimasi dengan *smoothing* parameter. Tujuan lainnya adalah memilih K yang tidak terlalu besar agar waktu perhitungan yang dibutuhkan tidak terlalu lama atau MSE yang lebih besar dari seharusnya. Rumus *fixed selection method* didefinisikan sebagai berikut:

$$K = \min(\tfrac{1}{4} \times \text{banyaknya } x_i \text{ yang } unique, 35) \quad (2.17)$$

Persamaan diatas merupakan metode yang umumnya digunakan untuk pemilihan jumlah knot dan penentuan lokasi knot yang optimum ditentukkan melalui kuantil ke-Kk dari x yang *unique*, dengan rumus sebagai berikut (Ruppert,et al., 2003)

$$K_k = \left(\frac{k+1}{K+2} \right), k = 1,2, ..., K \quad (2.18)$$

2.6 Pendugaan Area Kecil dengan Pendekatan Semiparametrik *Penalized Spline*

Pendugaan area kecil (SAE) adalah pendekatan yang digunakan untuk mengungkapkan hubungan antara variabel *interest* dengan variabel pendukung sebagai model linear dengan tambahan pengaruh acak area kecil. Dimisalkan θ merupakan vektor dari parameter *small area* yang berukuran $m \times 1$ dan diasumsikan vektor tersebut merupakan estimator langsung $\hat{\theta}$. Jika dinyatakan $m \times q$ adalah matriks dari variabel penyerta dari level area $x_i = \left(x_{1i}, x_{2i}, ..., x_{pi}\right)^T$

sehingga model SAE berbasis area dapat ditulis seperti persamaan (2.4) adalah sebagai berikut:

$\boldsymbol{\theta_i} = \boldsymbol{x}_i^T\ \boldsymbol{\alpha} + \boldsymbol{b_i v_i} + \boldsymbol{e_i}$; $\quad i = 1,2, ..., m$; $v_i \sim N(0, \sigma_v^2)$

Dimana b_i merupkan konstanta positif yang diketahui, v_i adalah pengaruh acak spesifik yang diasumsikan memiliki distribusi normal $v_i \sim N(0, \sigma_v^2)$. Menurut Giusti et al (2012), model SAE berbasis area berbasis area ini menghasilkan estimasi *small area* yang terpercaya dengan mengkombinasikan model SAE dan dan model regresi yang meminjam kekuatan dari domain lain, ketika asumsi ini tidak terpenuhi model SAE level area menyebabkan estimator bias dari parameter daerah kecil. Spesifikasi semiparametrik dari model SAE yang memungkinkan yaitu adanya hubungan nonliniear antara $\widehat{\boldsymbol{\theta}}$ dan variabel penyerta $\boldsymbol{X}$, dapat diperoleh menggunakan pendekatan *penalized spline.*

Seperti pada persamaan (2.8), model semiparametrik dengan satu respon x_1 dapat ditulis $\tilde{m}(x_1)$ dimana fungsi dari $\tilde{m}(.)$ tidak diketahui akan tetapi diasumsikan cukup baik sehingga diberikan fungsi spline adalah sebagai berikut:

$$m(x_i) = \beta_0 + \beta_i x_i + \cdots + \beta_p x_1^p + \sum_{j=1}^{k} \gamma_j (x_i - k_j)_+^p,$$

Dengan p adalah derajat spline, $(x_i - k_j)_+ = maks\{0, (x_i - k_j)\}, k_j$ dimana j = 1, ... ,K merupakan himpunan titik knot. Dengan menetapkan $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ adalah $(p + 1)$ vektor koefisien fungsi polinomial, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T$ adalah vektor koefisien *spline*,

Dengan $(x_i - k_j)_+^p \begin{cases} = (x_i - k_j)_+^p & untuk\ x_i \geq k_j \\ = 0 & untuk\ x_i < k_j \end{cases}$

Menurut opsomer, et al (2008) model tersebut diindikasikan akan *overparameterized* sehingga akan menyebabkan *overfitiing* untuk menghindari hal tersebut ditambahkan penalty pada parameter *spline* dengan meminimumkan fungsi *Penalized Least Square* sehingga didapatkan hasil sesuai dengan persamaan (2.15).

Pada penelitian ini pendekatan SAE dengan menggunakn *penalized spline* sebagai efek random menghasilkan

$y_i = m(\boldsymbol{X_i}; \boldsymbol{\beta}) + \boldsymbol{\varepsilon_i}$

$= \boldsymbol{X_i} * \boldsymbol{\beta} + \boldsymbol{\varepsilon_i}$

$= \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{\gamma} + \boldsymbol{\varepsilon_i}$ (2.19)

Dimana: $\boldsymbol{X}_i \boldsymbol{\beta} = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_+^p$ (komponen parametrik yang merupakan *fixed componen*); $\boldsymbol{Z}_i \boldsymbol{\gamma} = Z_{1i}\gamma_1 + \cdots + Z_{ki}\gamma_k = ((X_i - K_1)_+^p \beta_{pi1} + \cdots + (X_i - K_1)_+^p \beta_{piK}$ (deviasi dari komponen parametrik dengan random efek); dan $\gamma = \gamma_1, \ldots, \gamma_{\gamma k})$ yang berasumsi mean 0 dan variansi σ_γ^2. Model *penalized spline* merupaka model *random efect* yang dikombinasikan dengan model SAE berbasis area agar mendapatkan estimasi pendugaan area kecil secara semiparametrik berdasarkan *linear mixed model*. Dari persamaan (2.3) dan persamaan (2.19) didapatkan model semiparametrik Fay-Herriot dapat ditulis sebagai berikut:

$$\hat{\theta} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X_1} \end{bmatrix} [\alpha, \beta] + \boldsymbol{Z}\gamma + \boldsymbol{b}v + e \quad (2.20)$$

Menurut Giusti et al (2012), jika terdapat variabel lain yang perlu disertakan dalam model, variabel tersebut dapat ditambahkan kedalam *X* sebagai matriks efek tetap. Opsomer (2004) menggunakan *penalized spline* untuk mengestimasi

area kecil dan menambahkan pengaruh acak kecil pada model sehingga didapatkan persamaan:

$$\hat{\theta} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + \boldsymbol{b}v + e \quad (2.21)$$

Persamaan di atas terdiri dari fungsi spline yang merupakan fungsi semiparametrik $\boldsymbol{X}\beta + \boldsymbol{Z}\gamma$ dan pengaruh acak area kecil ($\boldsymbol{b}v$). Nilai estimasi pada $\hat{\beta}$ semiparametrik *penalized spline* untuk penduga area kecil dengan menggunakan (MLE) sehingga didapatkan

$$\hat{\beta} = (\boldsymbol{X^T V^{-1} X})^{-1} \boldsymbol{X^T V^{-1} Y}$$

Jika komponen varians tidak diketahui, maka setelah estimator β dan prediktor γ diperoleh, Estimasi komponen varians berdasarkan ML bias, maka digunakan metode REML (*Restricted Maximum Likelihood*).

2.7 *Best Linier Unbiased Prediction* (BLUP) dan *Empirical Best Linier Unbiased Prediction* (EBLUP)

Model *small area* terbagi menjadi model area level dan model unit level. Metode BLUP dan EBLUP salah satu metode yang digunakan untuk meminimumkan MSE. Pada metode BLUP, variansi pengaruh acak diasumsikan telah diketahui. Sedangkan pada metode EBLUP nilai variansi pengaruh acak *small area* tidak diketahui sehingga harus ditaksir dengan menggunakan metode *Maximum Likelihood* (ML)

Misalkan data memenuhi model linear campuran berikut:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{e} \quad (2.22)$$

Dimana

$\boldsymbol{y}$ adalah vektor data observasi berukuran $n \times 1$

$\boldsymbol{X}$ dan $\boldsymbol{Z}$ adalah matriks berukuran $n \times p$ dan $n \times$h yang diketahui

$\boldsymbol{\gamma}$ dan $\boldsymbol{e}$ adalah berdistribusi saling bebas dengan rataan 0 dan ragam $\boldsymbol{G}$ dan $\boldsymbol{R}$ yang tergantung pada parameter $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q)^T$, diasumsikan bahwa $\boldsymbol{\delta}$ adalah himpunan bagian dari ruang *Euclidean* sedemikian sehingga:

$$\boldsymbol{Var(y) = V = V(\delta) = R + ZGZ^T}$$

Adalah non singular untuk semua δ yang terdapat dalam himpunan bagian tersebut, dimana $Var(\boldsymbol{y})$ adalah matriks varians covarians dari $\boldsymbol{y}$.

Parameter yang akan diduga merupakan kombinasi linear $\boldsymbol{\mu = 1_l^T \beta + m^T v}$ (Rao,2003). Vektor $\mathbf{1}$ dan $\boldsymbol{m}$ adalah konstan. Penduga linear dari μ adalah $\boldsymbol{\hat{\mu} = \alpha^T \beta + b}$ untuk $\boldsymbol{a}$ dan $\boldsymbol{b}$ diketahui. Sehingga penduga tak bias $\boldsymbol{\mu}$

$E(\hat{\mu}) = E(\mu)$

E adalh ekpektasi, MSE $\hat{\mu}$ didefenisikan sebagai MSE $(\hat{\mu}) = E(\hat{\mu} - \mu)^2$

Jika $\hat{\mu}$ adalah penduga tak bias dari μ, maka MSE $(\hat{\mu}) = E(\hat{\mu} - \mu)^2 = Var\,(\hat{\mu} - \mu)^2$

Estimator BLUP μ dan δ diketahui sebagai berikut:

$$\boldsymbol{\hat{\mu}^H = t(\delta, y) = (I^T \tilde{\beta} + m^T G Z^T V^{-1} (y - X\tilde{\beta})} \quad (2.23)$$

Dimana

$$\boldsymbol{\tilde{\beta} = \tilde{\beta}(\delta) = GZ^T V^{-1} (y - X\tilde{\beta})} \quad (2.24)$$

Merupakan *best linear unbiased* estimator (BLUE) dan β dan

$$\boldsymbol{\tilde{v} = \tilde{v}(\delta) = GZ^T V^{-1} (y - X\tilde{\beta})} \quad (2.25)$$

Keterangan H pada $\tilde{\mu}$ adalah Henderson yang mengusulkan persamaan (2.23)

Penduga BLUP tergantung pada ragam δ yang biasanya tidak diketahui. Jika δ diduga dengan $\tilde{\delta}= \tilde{\delta}(y)$, maka akan diperoleh *Empirical Best Linear Unbiased Prediction* (EBLUP) yang tetap merupakan penduga tak bias bagi μ. Penduga δ diperoleh melalui metode ML atau REML (Rumiati, 2012).

2.8 Pendugaan MSE dengan menggunakan Metode *Jackknife* dan Pendugaan MSE Tidak Langsung

Menurut Baillo dan Molina (2009), tujuan dari prosedur dan teknik yang digunakan dalam SAE adalah untuk memperoleh estimasi dengan tingkat presisi yang tinggi pada area kecil tersebut. Tingkat presisi estimator ini dapat digambarkan oleh *Mean Square Error* (MSE). Penerapan *jackknife* pada SAE dilakukan untuk mengkoreksi pendugaan MSE.

Fay dan Herriot (1979) mengembangkan model $y_i = x_i^T \beta + v_i + e_i$ sebagai dasar dalam pengembangan SAE. Untuk selanjutnya diasumsikan bahwa β dan σ_v^2 tidakdiketahui, akan tetapi σ_{e1}^2 diketahui, dengan $\beta_i = \sigma_{ei}^2/(\sigma_v^2 + \sigma_{ei}^2)$ maka

$$MSE\left(\hat{\theta}_i^{EBLUP}\right) = E\left(\hat{\theta}_i^{EBLUP} - \theta_i\right)^2$$

$$= Var\left(\hat{\theta}_i^{EBLUP}\right) + \left(bias\left(\hat{\theta}_i^{EBLUP}\right)\right)^2$$

Persamaan tersebut dapat diuraikan menjadi

$$MSE\left(\hat{\theta}_i^{EBLUP}\right) = MSE\left(\hat{\theta}_i^{EBLUP}\right) + E(\hat{\theta}_i^{EBLUP} - \hat{\theta}_i^{BLUP})^2 \qquad (2.26)$$

Metode *jackknife* pertama kali diperkenalkan oleh tukey pada tahun 1958 dan kemudian berkembang sebagai suatu metode untuk mengoreksi bias pada suatu estimator. Dengan melakukan penghapusan terhadap observasi ke-i untuk i

$=1,2, \ldots, m$ dan kemudian dilakukan pendugaan parameter misal $\hat{\theta}$, maka penduga bias diduga dengan

$$bias(\hat{\theta}) = (m-1)[\hat{\theta}_{(i)} - \hat{\theta}]$$

Dengan $\hat{\theta}_{(.)} = m^{-1} \sum_i^m \hat{\theta}_{(.)}$

Penduga *jackknife* diperoleh dari

$$\hat{\theta}_{jack} = \hat{\theta} - bias(\hat{\theta}) \text{ dan } v_{jack}(\hat{\theta}) = \frac{(n-1)}{n} \sum_i^m [\hat{\theta}_{(i)} - \hat{\theta}]^2$$

Penerapan *jackknife* pada SAE dilakukan untuk mengkoreksi pendugaan MSE akibat adanya pendugaan α dan σ_v^2. Persamaan (2.24) setara dengan $g_{1i}(\sigma_i^2) + (bias)^2$ jika σ_i^2 diduga.

Dengan *u* adalah replikasi *jackknife* dan *i* adalah banyaknya data, maka prosedur *jackknife* $MSE(\hat{\theta}_i^{EBLUP})$ pendugaan tidak langsung berdasarkan persamaan (2.26) adalah sebagai berikut:

1. $MSE(\hat{\theta}_i^{EBLUP})$ diketahui oleh:

   $$MSE_i(\hat{\theta}_i) = h_{1i} + h_{2i}$$

2. Penduga variansi $MSE(\hat{\theta}_i^{EBLUP})$ dengan menghitung:

   $$h_{1i} = g_{1i}(s_v^2) - \left(\frac{m-1}{n}\right) \sum_{u=1}^{m} [g_{1i}(s_{v(-u)}^2) - g_{1i}(s_v^2)]$$

   Dimana $g_{1i}(s_{v(-u)}^2)$ diperoleh dengan menghapus pengamatan ke-*u* pada himpunan data $g_{1i}(s_v^2)$ dan *u* = 1,2, ... , m.

   $$s_v^2 = (m-1)^{-1} \sum_i (y_i - \hat{y})^2 - \sigma_e^2$$

   $$s_{v(-u)}^2 = (m-2)^{-1} \sum_{i(-u)} (y_i - \hat{y})^2 - \sigma_e^2$$

3. Menduga $E(\hat{\theta}_i^{EBLUP} - \hat{\theta}_i^{BLUP})^2$ dengan menghitung:

$$h_{1i} = \left(\frac{m-1}{n}\right)\sum_{u=1}^{m}\left[\left(\hat{\theta}_{i(-u)}\right) - \left(\hat{\theta}_i\right)\right]^2$$

Dimana $\left(\hat{\theta}_{i(-u)}\right)$ diperoleh dengan menghapus pengamatan ke-u pada himpunan data $\left(\hat{\theta}_i\right)$

Nilai MSE pendugaan langsung dengan u adalah banyak replikasi *jackknife* dan i adalah banyak data, maka prosedur *jackknife* pendugaan langsung berdasarkan persamaan (2.21) adalah sebagai berikut:

1. MSE pendugaan langsung didekati oleh:

   $MSE_i(y_i) = h_{1i} + h_{2i}$

2. Menduga variasi MSE pendugaan langsung dengan menghitung:

   $h_{1i} = g_{1i}(s_v^2) - \left(\frac{m-1}{n}\right)\sum_{u=1}^{m}\left[g_{1i}\left(s_{v(-u)}^2\right) - g_{1i}(s_v^2)\right]$

   Dimana $g_{1i}\left(s_{v(-u)}^2\right)$ diperoleh dengan menghapus pengamatan ke-u pada himpunan data $g_{1i}(s_v^2)$ dan u = 1,2, … , m. Dengan:

   $s_v^2 = (m-1)^{-1}\sum_i (y_i - \hat{y})^2 - \sigma_e^2$

   $s_{v(-u)}^2 = (m-2)^{-1}\sum_{i(-u)} (y_i - \hat{y})^2 - \sigma_e^2$

3. Menduga nilai h_{2i} dengan menghitung:

   $$h_{2i} = \left(\frac{m-1}{n}\right)\sum_{u=1}^{m}\left[\left(y_{i(-u)}\right) - \left(y_i\right)\right]^2$$

   Dimana $\left(y_{i(-u)}\right)$ diperoleh dengan menghapus pengamatan ke-u pada himpunan data (y_i). Nilai RMSE diperoleh setelah mendapatkan nilai MSE melalui persamaan (2.26). Dimana *Root Mean Square Error* (RMSE) merpakan untuk mencari kesalahan dari rata-rata error pada observasi (Willmott dan

Matsuura 2005). RMSE dapat digunakan mencari tahu seberapa besar kesalahan pada data dari data model yang digunakan. RMSE dapat dijadikan sebagai indikator ketidakcocokan dalam pemodelan. RMSE dapat dicari dengan menggunakan:

$$RMSE(\hat{\theta}_i) = \frac{\sqrt{MSE(\hat{\theta}_i)}}{\hat{\theta}_i} \times 100\% \quad (2.27)$$