Data Mining

GEMASTIK 12

Analisis Kemunculan Titik Api menggunakan Algoritma Expectation Maximization dan Autoregressive Integrated Moving Average (ARIMA) di Provinsi Riau

Disusun oleh :

Ahmad Shohibus Sulthoni

Nur Azizah Harun

EDM22

120410575421

Fakultas Rekayasa Industri

Universitas Telkom

2019

DAFTAR ISI

1. Latar belakang

Titik panas adalah area vulkanik yang dihasilkan dari mantel yang secara anomali lebih panas dibandingkan mantel di sekitarnya. Titik panas bisa berada dekat maupun jauh dari batas-batas tektonik lempeng. Keberadaan titik panas menjadi salah satu faktor terjadinya kebakaran hutan dan lahan (karhutla).

Berdasarkan informasi dari Badan Meteorologi Klimatologi dan Geofisika (BMKG) yang mendeteksi sekitar 1.211 titik panas berada pada wilayah Sumatera. Khusus di Riau, 258 titik panas paling banyak di Kabupaten Indragiri Hilir (Inhil) yakni mencapai 143 titik. Kemudian di Kabupaten Pelalawan 47 titik, Indragiri Hulu (Inhu) 25 titik, Rokan Hilir (Rohil) 23 titik, Bengkalis 9 titik, Kuantan Singingi 3 titik, Rokan Hulu 2 titik, dan Kota Dumai ada satu titik. Titik-titik panas ini menjadi salah satu indikasi awal terjadinya karhutla yang memiliki dampak terhadap kegiatan masyarakat.

Hal ini harus segera ditanggulangi, salah satunya dengan cara memberikan informasi kepada masyarakat terkait informasi tentang daerah yang berpotensi menjadi titik panas agar dapat lebih peduli. Data terkait titik panas beredar sangat banyak sehingga tidak mungkin mendapatkan kesimpulan satu per satu dari data tersebut. Oleh karena itu, diperlukan sebuah metode khusus untuk mengolah data tersebut.

Dalam studi ilmu komputer, terdapat bidang yang khusus mempelajari mengenai pengelolaan terhadap data dalam jumlah yang banyak, yaitu *data mining*. Teknik-teknik dalam *data mining* memungkinkan pengolahan data yang banyak secara optimal, sehingga informasi-informasi yang bermanfaat dapat diperoleh dari data tersebut. Salah satu contoh pengaplikasian *data mining* yang akan kami bahas dalam makalah ini adalah pengolahan terhadap *dataset* titik panas pada provinsi Riau pada 1 Januari 2014 hingga 31 Agustus 2019.

2. Tujuan dan manfaat

Analisis yang kami lakukan terhadap data yang kami peroleh bertujuan untuk mencari menganalisis fenomena serta anomali data terhadap informasi titik panas yang berada di provinsi Riau.

Manfaat yang dapat diperoleh dari pembuatan makalah ini diharapkan dapat menghasilkan nilai yang optimal untuk penentuan daerah rawan titik panas yang berada pada Provinsi Riau. Analisis ini akan sangat berguna apabila dilakukan *stream data* baru secara terus menerus perbulannya, agar hasil analisis ini dapat menghasilkan informasi terkait daerah-daerah yang rawan akan kemunculan titik panas sehingga dapat membantu menanggulangi penyebab terjadinya kebakaran hutan dan lahan sejak dini.

3. Batasan yang digunakan

Pada penelitian ini, kami melakukan pengolahan data Provinsi Riau. Riau dipilih karena saat ini menjadi salah satu daerah yang mengalami dampak yang besar terkait karhutla yang terjadi, di sisi lain kami mengharapkan kedepannya kejadian karhutla tidak meluas di wilayah Indonesia lainnya. Karena data yang harus diproses relatif banyak, algoritma yang digunakan adalah algoritma-algoritma yang mampu berjalan cukup cepat, seperti Expectation Maximization dan Autoregressive Integrated Moving Average (ARIMA) .

4. Metode Penambangan Data

4.1.Perangkat Lunak

*4.1.1.*Google Fusion Tables

Google Fusion merupakan suatu aplikasi berbasis web, yang digunakan untuk mengumpulkan, serta visualisasi data dengan tujuan penggunaan bersama. *Google Fusion Tables* dapat membuat visualisasi berbentuk peta dengan bantuan Google Maps dan aplikasi ini digunakan dalam pembuatan *heatmap*. Kelebihan dari penggunaan *Google Fusion Tables* adalah hasil dari visualisasi yang dihasilkan cukup jelas dan dapat disesuaikan dengan kebutuhan saat mengolah data pada analisis yang kami lakukan.

4.1.2.Python (Scikit-Learn, matplotlib, statsmodels)

Python dipilih karena kemudahan implementasinya dan dapat dikembangkan dengan cepat. Python pada kasus ini digunakan untuk training data yang kami dapatkan, dan mengevaluasi keseluruhan data dengan cara menguji model algoritma yang diberikan pada data. Scikit digunakan untuk melakukan *preprocessing* data dan *clustering*. Matplotlib digunakan untuk visualisasi data berupa plot dan scatter. Statsmodels kami gunakan untuk pemodelan ARIMA.

4.2.Dataset

Dataset yang kami gunakan adalah data titik panas yang berada pada Provinsi Riau yang tercatat mulai dari 1 Januari 2014 hingga 31 Agustus 2019. Terdapat 188.148,9 titik panas yang diperoleh dari dataset yang kami gunakan.

Dataset yang kami gunakan memiliki 9 atribut, yaitu Lintang (deg), Bujur (deg), Tanggal (dd/mm/yyyy), Waktu Akuisisi (UTC), Tingkat Kepercayaan (%), Satelit, Kecamatan, Kabupaten dan Provinsi. Berikut penjelasan untuk masing-masing atribut yang kami gunakan :

1. Lintang : Sebagai latitude untuk sistem koordinat geografis permukaan bumi, yang menentukan sebuah lokasi di muka bumi.
2. Bujur : Sebagai longitude untuk sistem koordinat geografis permukaan bumi, yang menentukan sebuah lokasi di muka bumi.
3. Tanggal : Menunjukkan hari pengambilan data terkait titik panas yang berada pada Provinsi Riau.
4. Waktu Akuisisi : Menunjukkan waktu pengambilan data terkait titik panas yang berada pada Provinsi Riau.
5. Tingkat Kepercayaan : Menunjukkan keyakinan sensor menerima informasi terkait titik panas.
6. Satelit : Menerima sensor terkait keberadaan titik panas.
7. Kecamatan : Menunjukkan lokasi spesifik mengenai letak titik panas.
8. Kabupaten : Menunjukkan lokasi spesifik mengenai letak titik panas.
9. Provinsi : Menunjukkan lokasi spesifik mengenai letak titik panas.

4.3.Algoritma

4.3.1 Gaussian Mixture Model

Model campuran Gaussian (GMM) adalah kategori model probabilistik yang menyatakan bahwa semua titik data yang dihasilkan berasal dari campuran distribusi Gaussian terbatas yang tidak memiliki parameter yang diketahui. Parameter untuk model campuran Gaussian diturunkan baik dari estimasi posteriori maksimum atau algoritma maksimalisasi ekspektasi berulang dari model sebelumnya yang terlatih dengan baik. Model campuran Gaussian sangat berguna dalam hal pemodelan data, terutama data yang berasal dari beberapa kelompok. [2]

Model matematika dari GMM adalah sebagai berikut :

$$\log\left(P(X,Z|\mu,\sigma,\pi)\right)=\sum_{i=1}^{n}\sum_{k=1}^{K}I(Z_i=k)\left(\log(\pi_k)+\log(N(x_i|\mu_k,\sigma_k))\right)$$

4.3.2 Expectation Maximization

Algoritma EM mencoba untuk menemukan estimasi kemungkinan maksimum untuk model dengan variabel laten. Pada bagian ini, kami menjelaskan pandangan yang lebih abstrak EM yang dapat diperluas ke model variabel laten lainnya.

Biarkan X menjadi seluruh rangkaian variabel yang diamati dan Z seluruh rangkaian variabel laten. Karena itu kemungkinan log:

$$\log\left(P(X|\Theta)\right) = \log\left(\sum_{Z} P(X, Z|\Theta)\right)$$

Gaussian Mixture Model via EM :

- **E-step**:

$$p_{ij}^{(m+1)} = \frac{G(y_i, \mu_j^{(m)}, \sigma_j^{(m)}) c_j^{(m)}}{\sum_{k=1}^{K} G(y_i, \mu_k^{(m)}, \sigma_k^{(m)}) c_k^{(m)}}$$

- **M-step**:

$$\begin{aligned}
\mu_j^{(m+1)} &= \frac{\sum_{i=1}^{n} y_i p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \\
(\sigma_j^{(m+1)})^2 &= \frac{\sum_{i=1}^{n} (y_i - \mu_j^{(m+1)})^2 p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \\
c_j^{(m+1)} &= \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(m+1)}
\end{aligned}$$

4.3.3 Autoregressive Integrated Moving Average

ARIMA sering juga disebut metode runtun waktu Box-Jenkins adalah model yang secara penuh mengabaikan independen variabel dalam membuat forecasting. ARIMA menggunakan nilai masa lalu dan sekarang dari variabel dependen untuk menghasilkan forecasting. ARIMA sangat baik ketepatannya untuk forecasting jangka pendek, sedangkan untuk forecasting jangka panjang ketepatan forecastingnya kurang baik (Syahrir, 2017).

Kami menggunakan ARIMA dengan pertimbangan time-series yang didapat tidak terlalu panjang. Artinya model ARIMA dapat berjalan dengan baik.

Model ARIMA non-seasonal diklasifikasikan sebagai model "ARIMA (p, d, q)", di mana :

- p adalah jumlah istilah autoregresif,
- d adalah jumlah perbedaan nonseasonal yang diperlukan untuk stasioneritas, dan
- q adalah jumlah kesalahan prakiraan keterlambatan dalam persamaan prediksi.

Persamaan peramalan umum dapat dikonstruksi menjadi:

$\hat{y}t = \mu + \phi 1\ yt\text{-}1 + \ldots + \phi p\ yt\text{-}p - \theta\ 1et\text{-}1 - \ldots - \theta\ qet\text{-}q$

4.4.Teknik

Dalam penambangan data ini, teknik-teknik yang digunakan antara lain:

4.4.1.Preprocessing

Sebelum dataset diproses, perlu dilakukan langkah-langkah tertentu agar dataset lebih mudah untuk diproses dan sesuai dengan kriteria yang diinginkan. Langkah-langkah preprocessing antara lain :

- Data Filtering

  Data Filtering adalah proses pembuangan data-data yang tidak memenuhi batasan yang ditentukan. Pembuangan data yang tidak memenuhi batasan dilakukan agar hasil penelitian relevan dengan batasan tersebut dan proses evaluasi dan analisis data menjadi lebih mudah dan cepat karena tidak ada data sampah yang terlibat dalam komputasi-komputasi yang dilakukan. Kami melakukan grouping data pada dataset input dan melakukan aggregation pada data yang telah di grouping.

- Data Standardization

  Standardisasi data termasuk mengatur kembali banyaknya baris atau kolom pada dataset dan mengubah nilai yang ada menjadi kisaran tertentu, misalnya data nominal dijadikan numerik, boolean, atau lainnya. Dalam penelitian ini, penulis melakukan standardisasi data, yaitu melakukan normalisasi pada data numerik

- Feature Extraction

  Feature extraction adalah pengurangan atribut pada dataset apabila ukuran dataset terlalu besar atau ada atribut yang berulang (redundant). Dalam penelitian ini, kami mengabaikan atribut waktu akuisisi dan satelit karena tidak relevan dengan tujuan analisis yang dilakukan.

4.4.2.Modeling

Setelah melalui tahap preprocessing, data akan dimodelkan untuk menggambarkan distribusi data tersebut dan hubungan antara data yang satu

dengan yang lain. Pada penelitian ini, teknik modeling yang digunakan antara lain:

- Clustering

  Clustering adalah teknik mengelompokkan data yang memiliki kemiripan karakteristik dan menjadikan data yang memiliki kesamaan karakteristik tersebut menjadi 1 kelompok menggunakan algoritma yang telah dijelaskan sebelumnya (GMM via EM).

- Regression Line Fitting

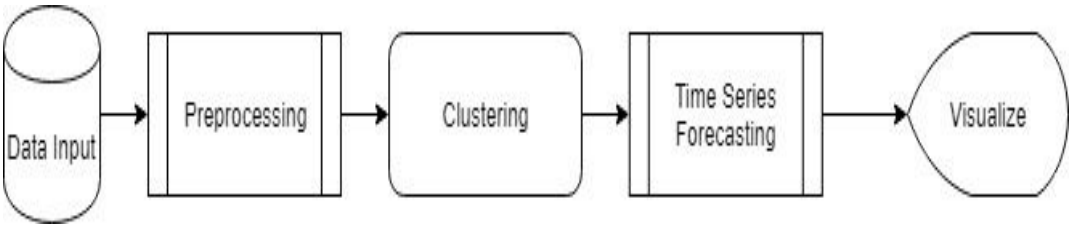
  Regression line fitting adalah teknik mengaproksimasi trend data dengan membentuk fungsi regresi yang mendekati distribusi data asli. Pada penelitian ini kami melakukan regression fitting menggunakan autoregressive method.

4.4.3.Inference

Inference adalah tahap penarikan kesimpulan pada hasil modeling yang telah dilakukan sebelumnya. Penarikan kesimpulan ini menggunakan konsep visualisasi matematis terhadap data output sehingga diperoleh informasi yang berkaitan dengan tujuan pengolahan data.

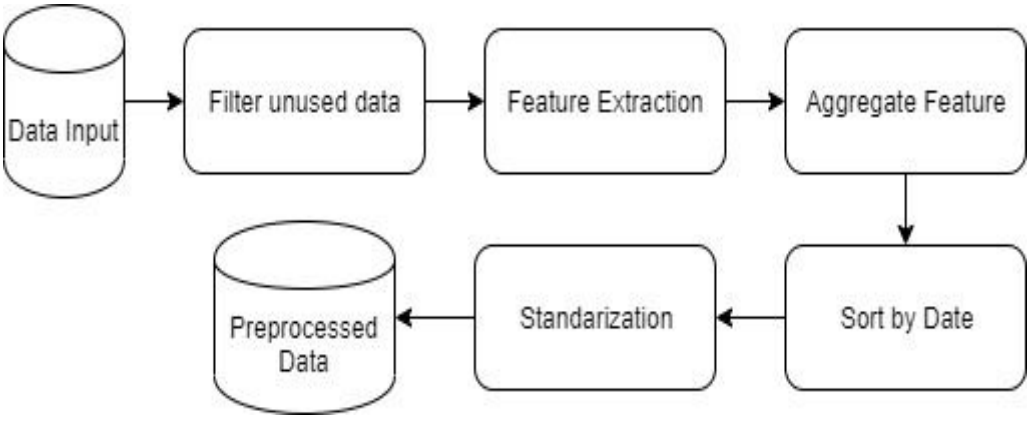
5. Desain dan implementasi Penambangan Data

5.1.Desain



Langkah yang akan dilakukan pertama kali adalah preprocessing sehingga data tersebut siap untuk melalui langkah selanjutnya pada pengolahan dataset ini. Langkah selanjutnya adalah clustering terhadap data tersebut. Kemudian dilakukan time series forecasting untuk memberikan gambaran prediksi terhadap trend data.

5.2.Preprocessing



5.2.1.Filter Unused Data

Sebelum data diproses lebih lanjut, langkah pertama yang harus dilakukan adalah menghapus data yang tidak sesuai. Seperti data yang kami hapus adalah data yang tidak terkait dengan keberadaan titik panas selain dari Provinsi Riau.

5.2.2.Feature Extraction

Agar pemrosesan data lebih mudah, beberapa atribut yang tidak sesuai dengan tujuan dari analisis kami hilangkan dari dataset. Atribut-atribut tersebut, antara lain

Agar data lebih mudah diproses, atribut-atribut yang tidak sesuai dengan tujuan dari analisis ini dihilangkan dari dataset. Atribut-atribut tersebut antara lain Waktu Akuisisi (UTC) dan Satelit.

5.2.3.Aggregate Feature

Fungsi agregasi sebagai penggabungan beberapa atribut untuk mendapatkan sebuah data lebih ringkas dan mudah untuk dipahami. Seperti yang kami lakukan dengan menggabungkan nilai pada atribut kecamatan dan kabupaten, sehingga diperoleh informasi terkait data dari satu kabupaten yang terdiri atas beberapa kecamatan.

5.2.4.Sort by Date

Fungsi mengurutkan data berdasarkan tanggal ataupun hari terjadinya akan memudahkan untuk melihat prediksi terjadinya pertumbuhan titik panas yang berada pada Provinsi Riau.

5.2.5.Standardisasi

Fungsi standarisasi digunakan untuk menghasilkan data yang memiliki informasi berupa pendistribusian secara merata terkait titik panas yang berada pada Provinsi Riau.

5.3.Input

5.3.1. Input Preprocess

Input terhadap preprocess yang kami lakukan menggunakan dataset asli berupa Comma Separated Value (CSV) dengan atribut antara lain Lintang (deg), Bujur (deg), Tanggal (dd/mm/yyyy), Waktu Akuisisi (UTC), Tingkat Kepercayaan (%), Satelit, Kecamatan, Kabupaten dan Provinsi.

5.3.2. Input Clustering

Input yang digunakan pada proses clustering merupakan dataset hasil preprocess yang telah dilakukan sebelumnya, yaitu CSV dengan atribut antara lain Lintang (deg), Bujur (deg), Tanggal (dd/mm/yyyy), Tingkat Kepercayaan (%), Kecamatan, Kabupaten dan Provinsi.

5.3.3. Input ARIMA

Data yang menjadi input ARIMA (Autoregressive Integrated Moving Average) adalah data yang sudah terdapat klaster masing-masing. Tujuan penulis melakukan ARIMA adalah untuk mengetahui prediksi klaster mana saja yang akan tumbuh dan mengalami peningkatan jumlah titik api. Data yang terlabel akan dilakukan streaming dengan karakteristik bulan.

5.4.Eksperimen

Eksperimen yang kami lakukan pada penambangan data ini adalah pemilihan parameter ARIMA (p,q,a). Karena karakteristik setiap cluster berbeda, maka parameter yang dipilih juga berbeda. Pemilihan parameter dilakukan pada rentang (5,1,0) , (3,1,0) , (1,1,0), (1,0,0).

Error didapatkan pada cluster 9,13 dan 19 ketika menggunakan parameter (5,1,0) karena klaster tersebut berisi data yang minim. Pada selain tiga klaster diatas, didapatkan hasil maksimal berupa plot prediction yang akan dijelaskan pada bab selanjutnya.

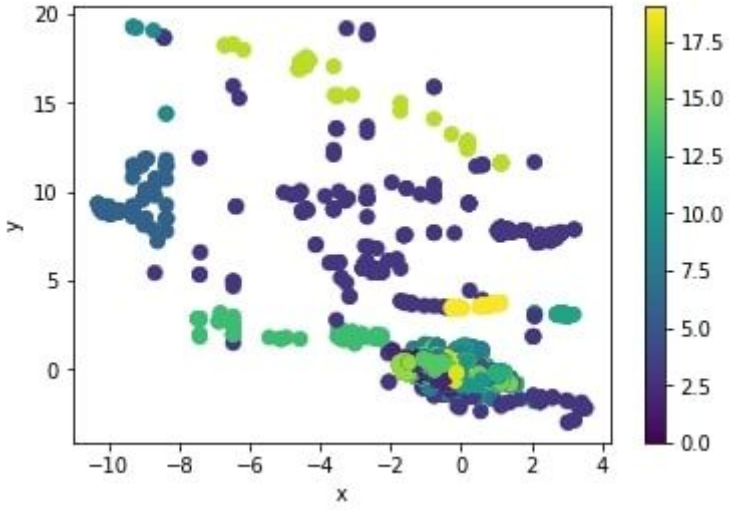
5.5.Output

5.5.1.Output preprocess

Output terhadap preprocess yang kami lakukan berupa dataset yaitu Comma Separated Value (CSV) dengan menghasilkan 7 atribut, antar lain Lintang (deg), Bujur (deg), Tanggal (dd/mm/yyyy), Tingkat Kepercayaan (%), Kecamatan, Kabupaten dan Provinsi.

5.5.2.Output Clustering

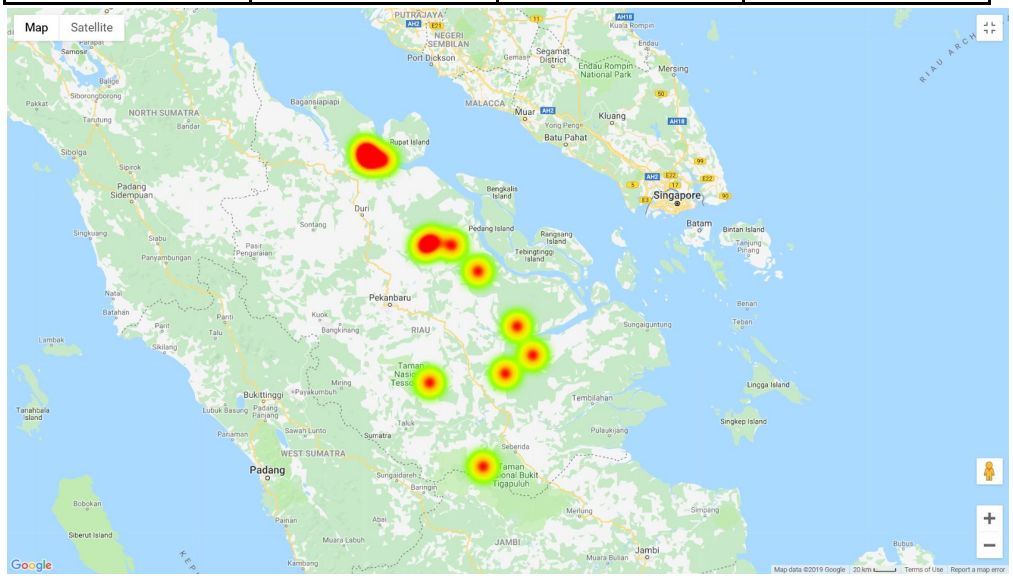
Output terhadap hasil clustering berupa pengelompokan isi dataset yang memiliki kemiripan data yang sama. Kemiripan dilihat melalui tiga features yaitu; lintang (deg), Bujur (deg) dan Tingkat Kepercayaan (%). Kami memilih 20 cluster dengan perkiraan persebaran tingkat kepercayaan yang merata. Dan didapatkan rentang 5 persen setiap cluster. Hasil dari clustering dapat dilihat di bawah ini:



Klaster	Lintang (deg)	Bujur (deg)	Area
0	1.003409	101.9655	Bandar Jaya, Siak Kecil, Bengkalis Regency, Riau
1	-0.0742	102.4243	Kerumutan, Pelalawan Regency, Riau
2	-0.14742	101.7776	Kesuma, Pangkalan Kuras, Pelalawan Regency, Riau
3	0.000157	111.8471	Linggam Permai, Kayan Hilir, Sintang Regency,
4	1.722594	101.2826	Lubuk Gaung,

			Sungai Sembilan, Dumai City, Riau 28826
5	1.722917	101.3898	P9FQ+5W Dumai, Dumai City, Riau
6	-9.21311	120.977	Savu Sea
7	0.999264	101.7405	Tasik Betung, Mandau River, Siak Regency, Riau
8	0.32222	102.5153	Teluk Binjai, Teluk Meranti, Pelalawan Regency, Riau
9	-8.30624	138.5601	Komolom, Kimaam, Merauke Regency, Papua
10	1.041652	101.8003	Tasik Betung, Mandau River, Siak Regency, Riau
11	3.737985	108.2529	P7Q3+55 Padang, South Cemaga, Natuna Regency, Riau Islands
12	1.7976	101.2426	Tanjung Penyembal, Sungai Sembilan, Dumai City, Riau
13	-4.06906	106.2039	Java Sea
14	0.078146	102.6533	Pulau Gelang, Kuala Cenaku, Indragiri Hulu Regency, Riau

15	1.754113	101.2539	Lubuk Gaung, Sungai Sembilan, Dumai City, Riau
16	-0.84809	102.2284	Alim, Batang Cenaku, Indragiri Hulu Regency, Riau
17	-1.16289	131.3247	Malaus, Salawati, Sorong, West Papua
18	0.789069	102.191	Paluh, Mempura, Siak Regency, Riau
19	1.41734	109.2142	Sekumbak, Lela, Tlk. Keramat, Kabupaten Sambas, Kalimantan Barat 79465



5.5.3.Output ARIMA

Output terhadap penggunaan algoritma ARIMA yang kami dapatkan berupa peramalan kemunculan titik api pada bulan selanjutnya. ARIMA yang kita lakukan menggambarkan trend perubahan titik api pada tiap klaster. Masing-masing klaster memiliki trend perubahan yang berbeda. Hasil ARIMA akan dijelaskan di bab selanjutnya. Trend ini dapat digunakan sebagai bentuk antisipasi kemunculan titik api pada masa yang akan datang. Klaster yang sangat perlu diwaspadai adalah pada klaster ke 5 dan ke 18.

Berikut adalah salah satu model results dari ARIMA :

ARIMA Model Results

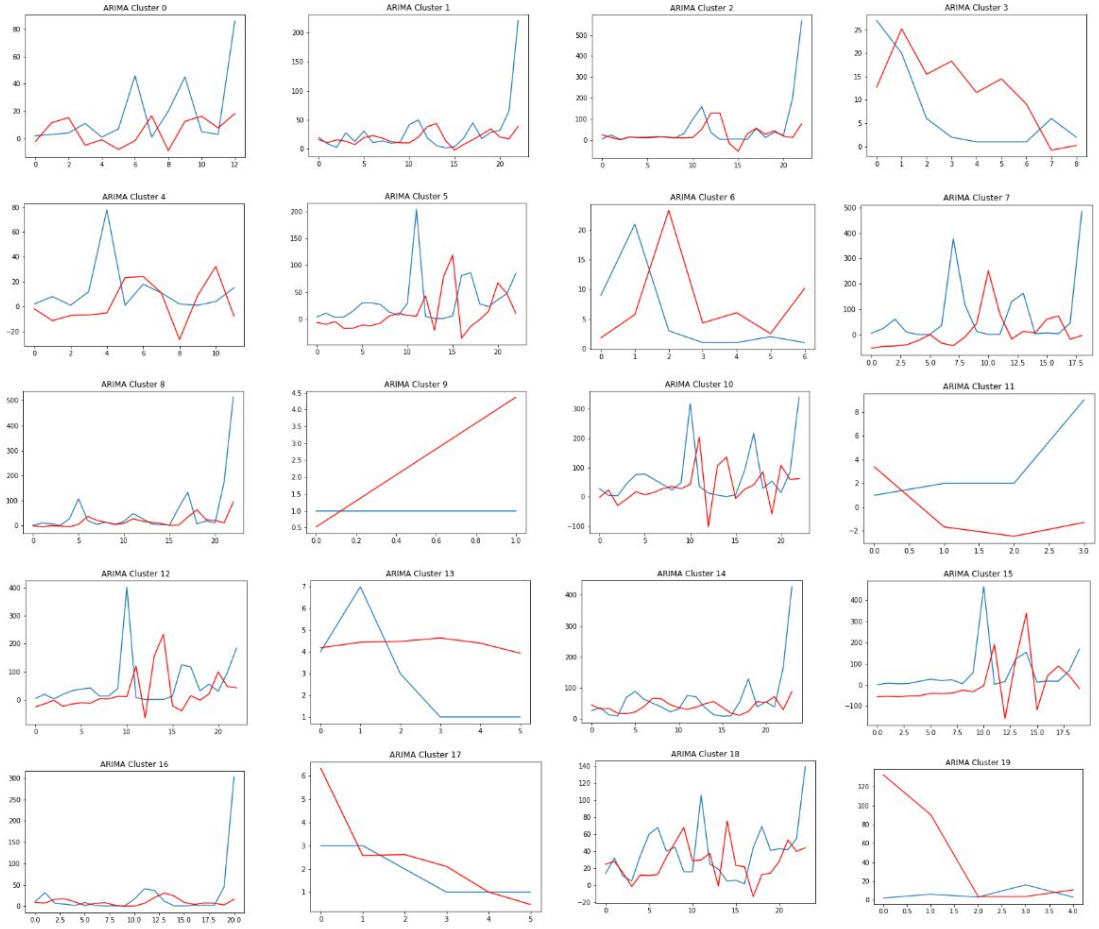
Dep. Variable:	D.count	No. Observations:	37
Model:	ARIMA(3, 1, 1)	Log Likelihood	-185.625
Method:	css-mle	S.D. of innovations	34.870
Date:	Thu, 26 Sep 2019	AIC	383.250
Time:	19:14:25	BIC	392.915
Sample:	1	HQIC	386.657

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5405	0.673	-0.803	0.428	-1.859	0.778
ar.L1.D.count	0.3185	0.180	1.768	0.087	-0.035	0.671
ar.L2.D.count	-0.2646	0.189	-1.401	0.171	-0.635	0.106
ar.L3.D.count	0.1471	0.330	0.445	0.659	-0.500	0.795
ma.L1.D.count	-1.0000	0.121	-8.232	0.000	-1.238	-0.762

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.2058	-1.7415j	1.7536	-0.2687
AR.2	-0.2058	+1.7415j	1.7536	0.2687
AR.3	2.2107	-0.0000j	2.2107	-0.0000
MA.1	1.0000	+0.0000j	1.0000	0.0000

6. Analisis



Terdapat 20 penggambaran plot dari hasil analisis yang kami lakukan. Setiap plot diagram mewakili setiap cluster, dimana garis berwarna biru menggambarkan data yang kami dapatkan dari dataset, sedangkan garis berwarna merah mewakili data dari hasil trainee data menggunakan algoritma ARIMA.

7. Kesimpulan

Pada penambangan data kali ini dapat diperoleh informasi mengenai trend kemunculan titik api di provinsi Riau. Tetapi ada beberapa kekurangan dalam proses klasterisasi, beberapa diantaranya adalah centroid dari beberapa cluster yang melenceng. Hal ini diakibatkan probability density yang ada dalam data mentah sebelum diolah tidak normal. Kemudian beberapa kendala dari penambangan data ini adalah istilah dari batan yang tidak diketahui oleh penulis. Beberapa feature yang dibuang bisa jadi merupakan feature penting seperti feature satelit.

8. Dokumentasi

Potongan source code Preprocessing :

In [5]:
```
useColumns = ['Lintang (deg)','Bujur (deg)','Tanggal (dd/mm/yyy)','Tingkat Kepercayaan (%)','Kecamatan', 'Kabupaten', 'Provinsi']
for a in data_riau_full.columns:
    if a not in useColumns:
        del data_riau_full[a]
data_riau_full.head()
```

Out[5]:

	Lintang (deg)	Bujur (deg)	Tanggal (dd/mm/yyy)	Tingkat Kepercayaan (%)	Kecamatan	Kabupaten	Provinsi
0	-0.139824	102.882179	2014-08-04	42.0	Gaunganakserka	Indragiri Hilir	Riau
1	1.983107	117.390663	2014-08-04	59.0	Gaunganakserka	Indragiri Hilir	Riau
2	1.979972	117.368645	2014-08-04	66.0	Gaunganakserka	Indragiri Hilir	Riau
3	-0.007676	102.744278	2014-08-04	25.0	Gaunganakserka	Indragiri Hilir	Riau
4	0.121050	102.805969	2014-08-04	16.0	Gaunganakserka	Indragiri Hilir	Riau

In [33]:
```
arrayKabupaten = data_riau_full.Kabupaten.unique()
print(len(arrayKabupaten))
print(arrayKabupaten)
```

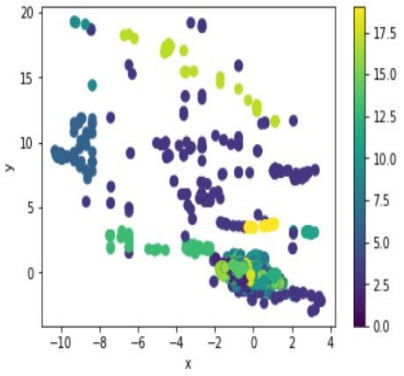
```
26
['Indragiri Hilir' 'Rokan Hulu' 'Bengkalis' 'Rokan Hilir' 'Kota Dumai'
 'Siak' 'Kampar' 'Indragiri Hulu' 'Indragiri Hulur' 'Pelalawan'
 'Kuantan Singingi' 'Kab. Indragiri Hulu' 'Kab. Siak' 'Kab. Rokan Hilir'
 'Kab. Bengkalis' 'Kab. Pelalawan' 'Kab. Kuantan Singingi'
 'Kab. Rokan Hulu' 'Kab. Kepulauan Meranti' 'Kab. Indragiri Hilir'
 'Kab. Kampar' 'Dumai' 'Kota Pekanbaru' 'Padang Lawas' 'Kepulauan Meranti'
 'Pekanbaru']
```

Potongan source code Clustering :

In [8]:
```
x = dfEMoutput['Lintang (deg)']
y = dfEMoutput['Bujur (deg)']
fig = plt.figure()
ax = fig.add_subplot(111)
scatter = ax.scatter(x,y,c=gmm_y,s=50)
# for i,j in center:
#     ax.scatter(i,j,s=50,c='red',marker='+')
ax.set_xlabel('x')
ax.set_ylabel('y')
plt.colorbar(scatter)

fig.show()
```

```
C:\Users\Ahmad Shohibus S\AppData\Roaming\Python\Python37\site-packages\matplotlib\figure.py:445: UserWarning: Matplotlib is currently using module://ipykernel.pylab.backend_inline, which is a non-GUI backend, so cannot show the figure.
  % get_backend())
```



Potongan source code ARIMA modeling :

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.arima_model import ARIMA
import datetime
from sklearn.metrics import mean_squared_error
from pandas.plotting import autocorrelation_plot

series = pd.read_csv('cluster_19_Arima.csv', usecols=['Tahun','count'],header=0, parse_dates=[0], index_col=0, squeeze=True)
X = series.values
size = int(len(X) * 0.66)
train, test = X[0:size], X[size:len(X)]
history = [x for x in train]
predictions = list()
for t in range(len(test)):
    model = ARIMA(history, order=(1,1,0))
    model_fit = model.fit(disp=0)
    output = model_fit.forecast()
    yhat = output[0]
    predictions.append(yhat)
    obs = test[t]
    history.append(obs)
    print('predicted=%f, expected=%f' % (yhat, obs))
error = mean_squared_error(test, predictions)
print('Test MSE: %.3f' % error)
# plot
plt.title("ARIMA Cluster 19")
plt.plot(test)
plt.plot(predictions, color='red')
plt.savefig("plot prediction/plot arima 19.png")
```

Daftar Pustaka

Slavia Putri, Athaya. *PENERAPAN MODEL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) UNTUK PREDIKSI KEMUNCULAN TITIK PANAS PADA KABUPATEN ROKAN HILIR*. Bandung. 2019

Amalia Khairani, Nabila. *PENERAPAN METODE CLUSTERING ALGORITMA K-MEANS SEBAGAI PENENTUAN DAERAH RAWAN TITIK API DI PROVINSI KALIMANTAN BARAT*. Bandung. 2019

https://www.statisticshowto.datasciencecentral.com/em-algorithm-expectation-maximization/

https://www.techopedia.com/definition/30331/gaussian-mixture-model-gmm

https://stephens999.github.io/fiveMinuteStats/intro_to_em.html