Remote-Ground based observations Merging Method For Visibility and Cloud Ceiling Assessment During the Night Using Data-Mining Algorithms

Driss BARI
Direction de la Météorologie Nationale
Casablanca, Morocco

Abdelhak Lemkhenter
Ecole Centrale SupElec
Paris, France

6 to 10 November 2017, Météo-France, Toulouse
Background & motivation

- Fog and Low cloud ceiling result in reduced horizontal and vertical visibility $\Rightarrow$ strong impact on aviation, marine and road traffic

- Derived satellite products have some limitations to distinct between fog and low clouds, particularly when they are obscured by other clouds (Cermak and Bendix, 2007)

- A nighttime product that depicts areas of the lowest cloud base heights by combining satellite data and surface temperature (Ellrod and Gultepe, 2007) $\Rightarrow$ under-detection usually occurs when thin cirrus overlies the stratus clouds
Background & motivation

- Data-mining methods have been used to cloud ceiling height assessment using both satellite data and NWP output. (*Bankert et al., 2004*)

Main objectives of this study

*To Evaluate the potential of Data mining methods:*

- To separately detect fog & low cloud ceiling
- To estimate visibility & cloud ceiling height
Study Domain & Synoptic Network

- 15 synoptic meteorological stations
- 6 stations, located in airports, are operating during the whole nighttime
Dataset & Study Period

- **Hourly Data**
- Study period: *winter months* (December, January, and February) from January 2014 to February 2017

- Meteosat Second Generation (MSG)
- All infrared channels at the four nearest pixels (over land) closest to each synoptic station
- Derived data:
  - Brightness Temperature
  - Brightness Difference Temperature

- Conventional meteorological parameters
- Lat-Lon coordinates and altitude of each synoptic station
**FOG**

- Horizontal visibility below 1000 m
- Visibility reduced below 1 km by other phenomena such as heavy rain are excluded

**LOW CLOUD CEILING**

- Cloudiness $\geq 4$ octas
- Cloud base height $\leq 1000$ m
Fog Frequency over the study period

For each synoptic station:

Frequency(%) = Number of foggy days during nighttime / Total available days

- Fog frequency over the region: 21% of all available days
- Per each station, Fog is often rare ( <= 5%) with a maximum of 29.6%
Low cloud ceiling frequency over the study period

- Low cloud ceiling frequency over the region: 64% of all available days over the region.

- Low cloud ceiling is frequent during the night in comparison with fog ( > 20% up to 64% )
Methodology: Tree Ensemble Approach

**XGBoost**: eXtreme Gradient Boosting for supervised learning

*(Chen and Guestrin, 2016)*

- Very widely used in machine learning competitions and industry.
- Invariant to scaling of inputs
- Regression tree ensemble defines how you make the prediction score. It can be used for classification, regression, ranking,...
Methodology: Tree Ensemble Approach

- We want **predictive** and **simple** models

\[
Obj(\Theta) = L(\Theta) + \Omega(\Theta)
\]

- Training Loss measures how well model fit on training data
- Regularization, measures complexity of model
Methodology : Tree Ensemble Approach

- **Model** : assuming we have $k$ trees

\[
\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}
\]

- **Parameters** :
  Instead of learning numerical weights, we are learning functions (trees)

Space of functions containing all Regression trees
Methodology: Tree Ensemble Approach

- How do we learn?
  Additive Training (Boosting)

Prediction of is sum of scores predicted by each of the tree

\[ f(\text{boy}) = 2 + 0.9 = 2.9 \]
\[ f(\text{old}) = -1 - 0.9 = -1.9 \]
Experiment set up

- A random split of all data: the available data (for all locations together) were randomly divided into training (75%) and testing (25%) sets.

- Equality between the numbers of occurrence and non-occurrence of the studied phenomenon into training and testing sets.

- The analysis was done for all locations as one.

- Tree boosting system: XGBoost (Chen and Guestrin, 2016).
Results
FOG and Low Cloud Ceiling Detection

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>FBIAS</th>
<th>PC</th>
<th>POD</th>
<th>FAR</th>
<th>TSS</th>
<th>ETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fog</td>
<td>1.01</td>
<td>0.83</td>
<td>0.83</td>
<td>0.17</td>
<td>0.66</td>
<td>0.49</td>
</tr>
<tr>
<td>Low cloud ceiling</td>
<td>1.02</td>
<td>0.71</td>
<td>0.72</td>
<td>0.30</td>
<td>0.41</td>
<td>0.26</td>
</tr>
</tbody>
</table>

- High performance for fog detection than for low cloud ceiling
- Frequency bias almost perfect
- PC and POD above 70% associated with low FAR (below 30%)
FOG Detection: Case study

Some uncertainties about spatial localization of fog!!

What is the potential of the unified model generalization over the study domain?
Spatial coverage
Frequency over the study period

Hourly Spatial Coverage over the Study Domain

- Fog
- Low Cloud Ceiling

Frequency (%)

Number of synoptic stations

1 2 3 >=4
### Generalization Error

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Fog</th>
<th>Low cloud ceiling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>POD</td>
<td>FAR</td>
</tr>
<tr>
<td><strong>WMO ID</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60120</td>
<td>1.00</td>
<td>0.17</td>
</tr>
<tr>
<td>60135</td>
<td>0.75</td>
<td>0.19</td>
</tr>
<tr>
<td>60136</td>
<td><strong>0.57</strong></td>
<td>0.14</td>
</tr>
<tr>
<td>60156</td>
<td>0.79</td>
<td>0.14</td>
</tr>
</tbody>
</table>

KDD-produced algorithms for fog and low cloud ceiling detection could be geographically dependent.
Visibility and Ceiling Height Estimation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Visibility</th>
<th></th>
<th></th>
<th>Cloud ceiling height</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
<td>CC</td>
<td>RMSE</td>
</tr>
<tr>
<td>Data source</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satellite</td>
<td>2659.25</td>
<td>1892.36</td>
<td>0.78</td>
<td>2448.38</td>
</tr>
<tr>
<td>Meteo</td>
<td>1170.57</td>
<td>675.24</td>
<td>0.96</td>
<td>1120.68</td>
</tr>
<tr>
<td>Satellite + Meteo</td>
<td>1167.81</td>
<td>676.92</td>
<td>0.96</td>
<td>1071.27</td>
</tr>
</tbody>
</table>

The performance of estimation algorithms is higher when local meteorological parameters are taking into account.
Conclusion

- High potential of KDD-produced algorithms for detecting fog than for low cloud ceiling

- High performance for estimating visibility and cloud ceiling height when local meteorological parameters are taken into account

- Performance of data mining methods could be geographically dependent
Perspectives

- **Input data**: enhancement of predictors such as NWP output

- **Data mining method**: tuning of XGBoost parameters, using other methods.

- **Predictability** of fog and low cloud ceiling using Data mining
Thank you