Universiteit Leiden
The Netherlands

Predicting damage due to rainfall

Christie Bavelaar (supervized by Jan N. van Rijn)

Leiden Institute of Advanced Computer Science

s.2155435@liacs.leidenuniv.nl

Research Question

How can machine learning techniques, combined with data from multiple sources be used to predict damage from rainfall in urban areas?

Twitter

To predict damage from rainfall we need some target attribute to approximate the reality. A list of twitter messages containing keywords related to damage from rainfall is used to create the target attribute. A twitter message results in a positive label for an example of rainfall in an area on the date the tweet has been send. When no twitter message has been send in that area on a given day that example has a negative label.

KNMI

The information about rainfall comes from the Royal Netherlands Meteorological Institute (KNMI). At the moment a python program adds all accounts of rain in a certain area during one day and saves it into a .csv file. In the future it would be possible to also add an attribute expressing the amount of rainfall during the 1 hour or 5 hours preceding the twitter message.

AHN

Filters

The information about terrain height is taken from the Actueel Hoogtebestand Nederland. In original form these are .tif images. It is possible to use filters on these images to preprocess the data. The sobol filter uses the derivative of the image to express the change in height. The gaussian filter is used image processing to blur and reduce noise. If and which filters would benefit the model is a question to be answered by conducting experiments.

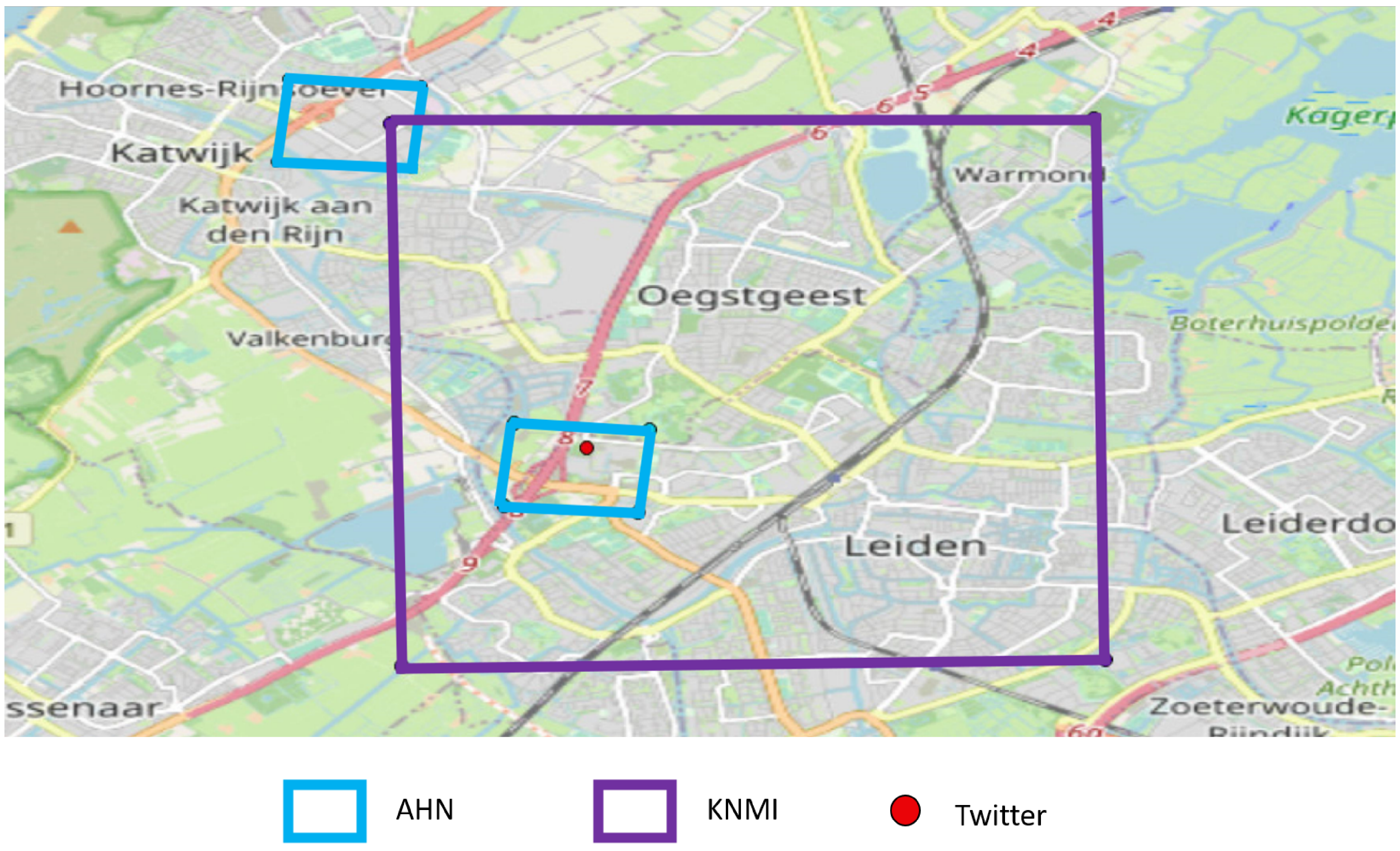
Choice of attributes

The height of the terrain around an example is incorporated into the model. With positive examples this is the location of the twitter message, in negative examples this is a random point in the area for which the rain attribute has been determined. A 20x20 grid of data points is taken around these coordinates. This results in each example having 400 attributes expressing the height of each point in the 20x20 grid. The choice of a 20x20 grid is an arbitrary one. Experiments can be used to determine an appropriate area to consider within the model.

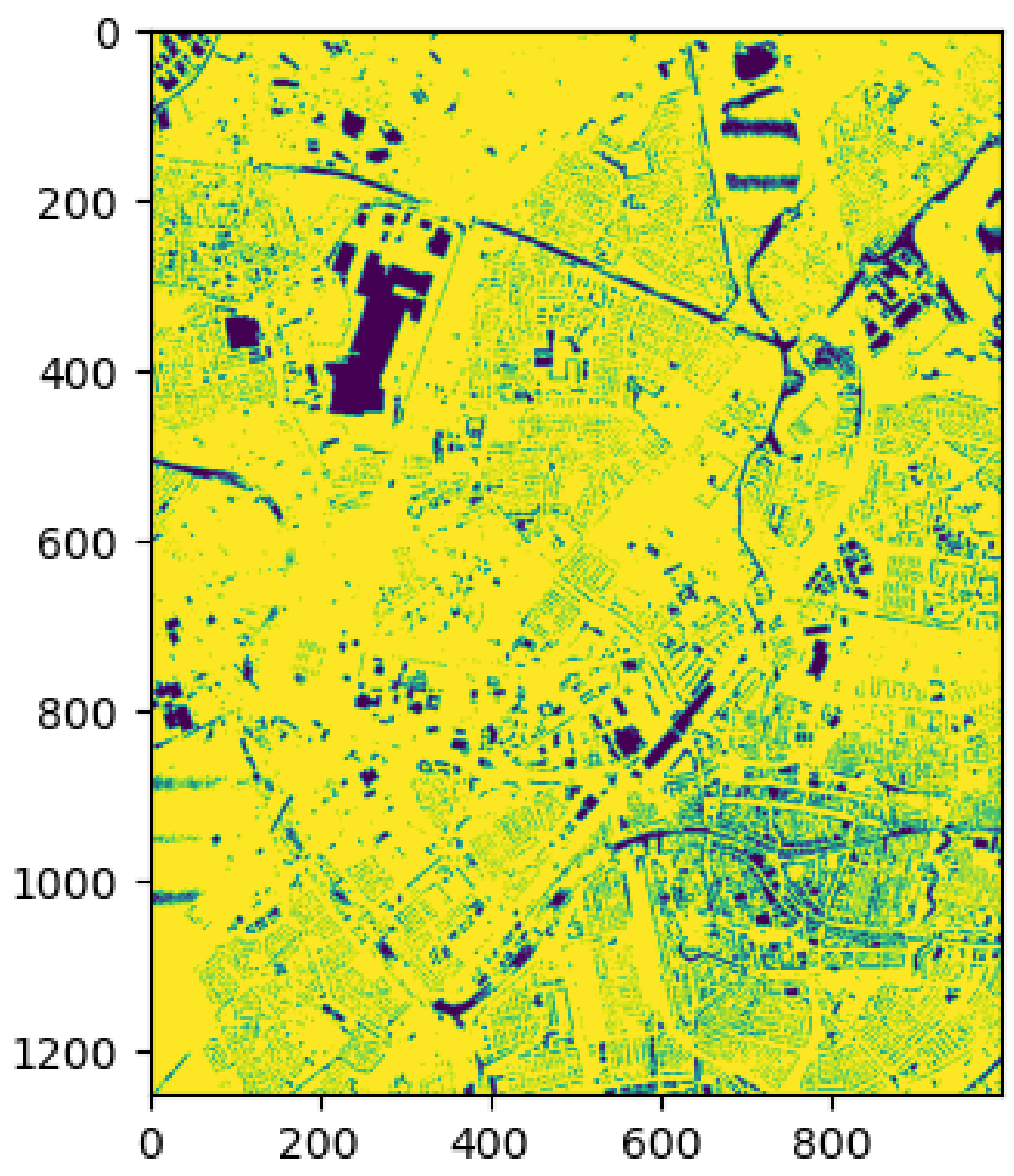
Urban areas

The research is specified for urban areas. The degree of urbanisation could have an effect on the accuracy of the model. By conducting experiments analysing the accuracy of the model when it is applied to a specific city this relationship can be explored further.

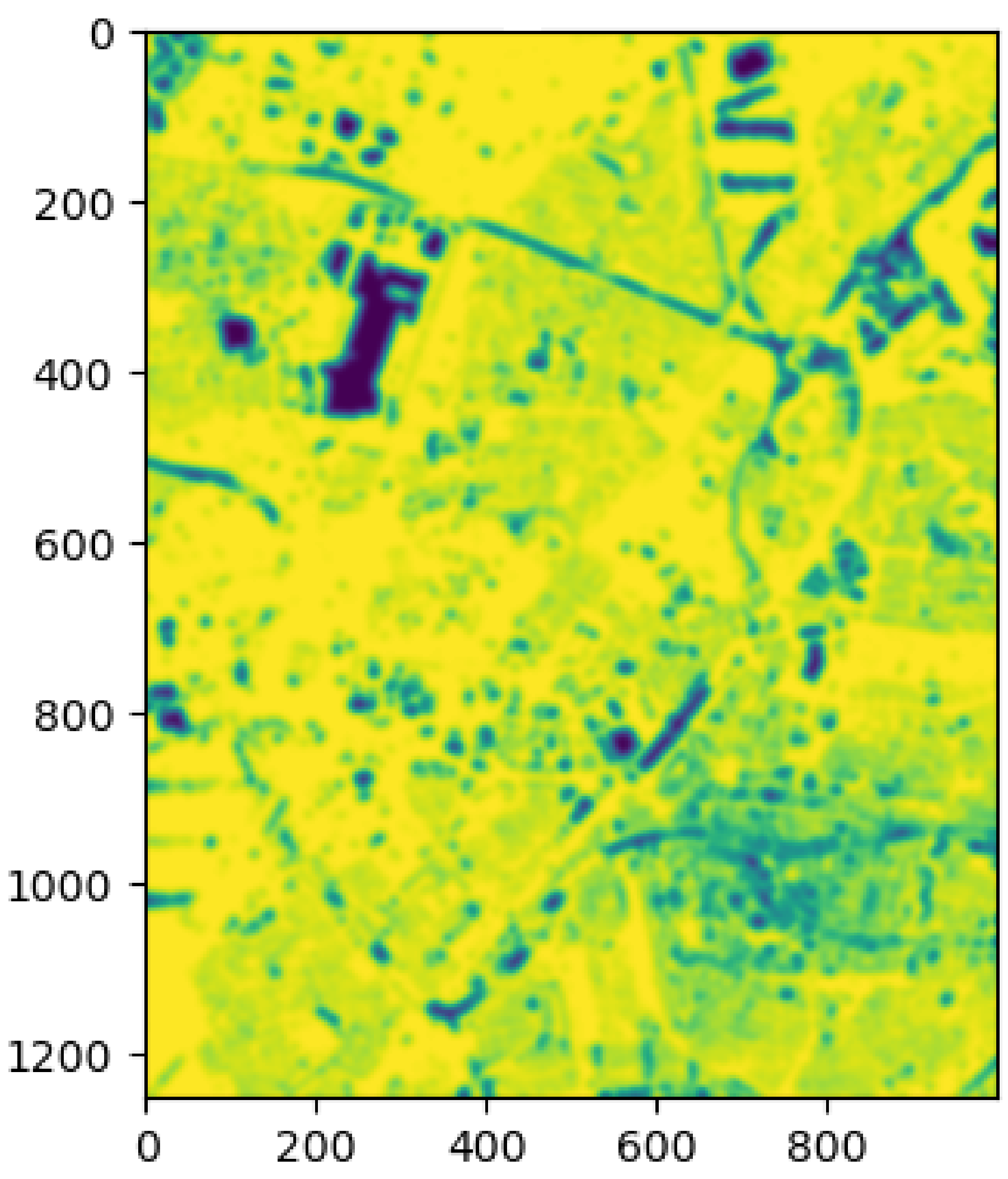
Data from multiple sources



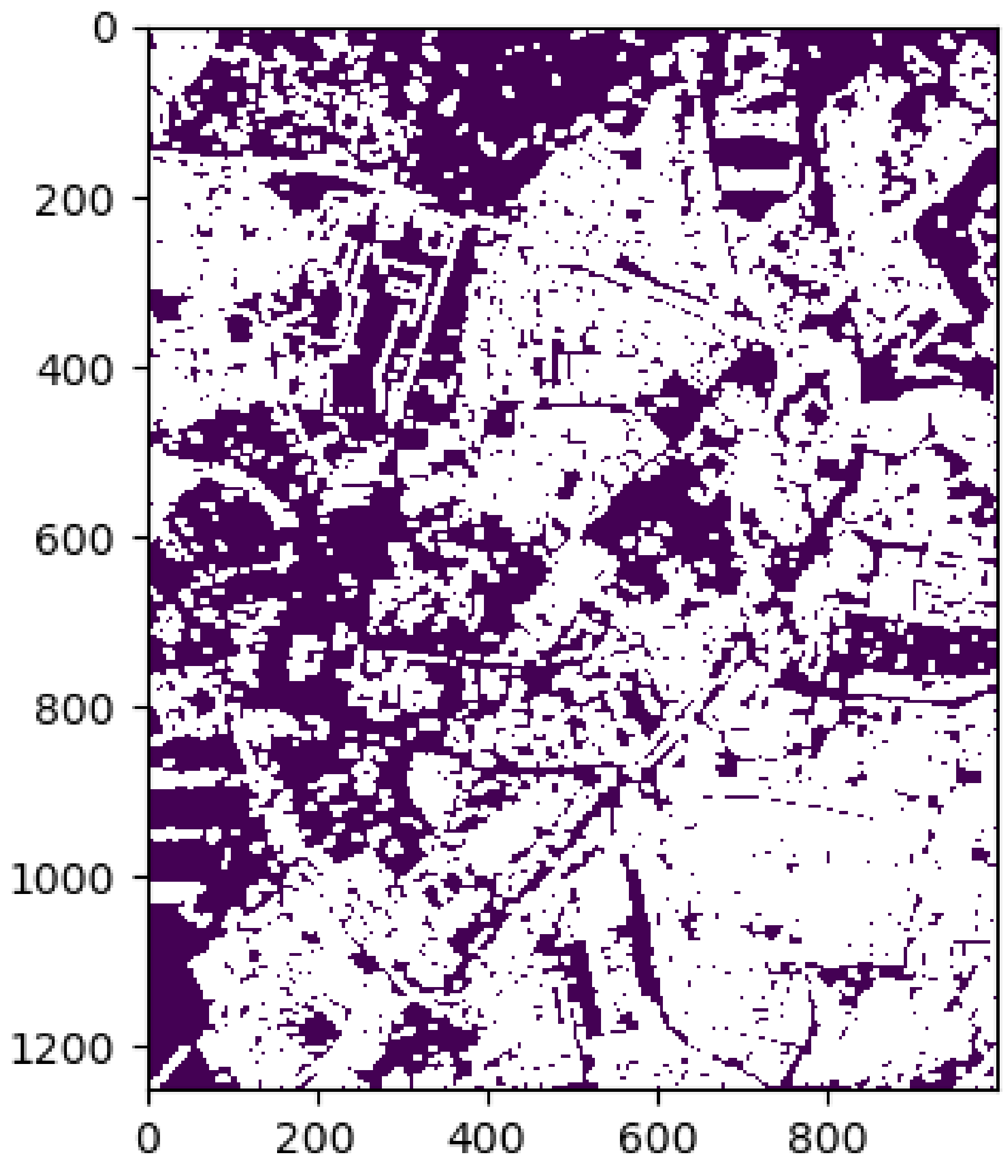
Original



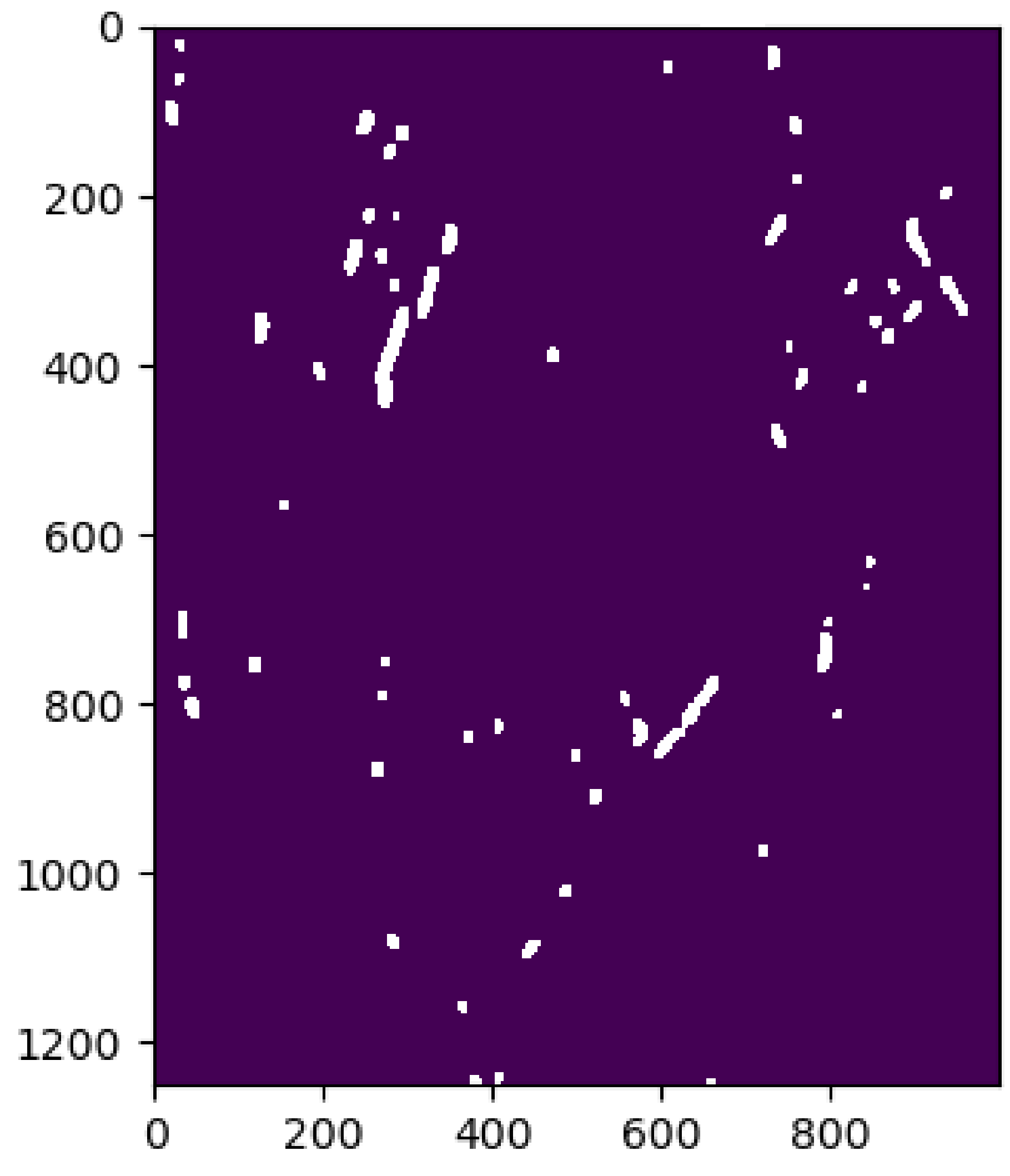
Gaussian filter



Sobol filter



Sobol & Gaussian filter


0
200
400
600
800
1000
1200
0
200
400
600
800