

# Predicting Ozone Peaks : A combined CBR and cell mapping approach

**David Pearson<sup>a</sup>, Mireille Batton-Hubert<sup>b</sup> and Gerardo Herrera Garcia<sup>b</sup>**

<sup>a</sup> EURISE (Roanne Research Group) University Jean Monnet, Saint-Etienne,  
IUT de Roanne, 42334 Roanne Cedex, France  
david.pearson@univ-st-etienne.fr

<sup>b</sup> Centre SITE Ecole Nationale Supérieure des Mines de Saint-Etienne,  
158 Cours Fauriel, 42023 Saint-Etienne Cedex 02, France  
batton@emse.fr

**Abstract:** In this paper we present a new approach for predicting ozone peaks when monitoring atmospheric pollution. Our main idea is that atmospheric pollution is closely monitored and over the past 5 to 10 years we have built up fairly extensive databases concerning it. Thus, rather than looking for a physical model in the first instance, maybe we should look for patterns and repeatability in the historical data. The approach presented is a hybrid one based on case based reasoning and cell mapping. Essentially, we forget all about a model in the first instance and simply search historical data to see if we can construct cases. We then extend this idea to include notions from cell mapping to see if we can build a cell map from the cases. The method is applied to real data coming from the Rhône-Alpes region of France.

**Keywords:** prediction of ozone peaks; atmospheric pollution; case based reasoning

## 1. INTRODUCTION

The problem of predicting ozone peaks is fairly well known in France and in certain other countries, notably situated around the Mediterranean area, suffering from the same phenomena. The problem details differ from region to region and country to country, but the overall objective remains the same. Basically, given known meteorological data and atmospheric pollution data up to a certain time today, we wish to predict the maximum level of ozone for tomorrow. The accuracy and the robustness of the method of prediction are extremely important because decisions are taken based on the predictions. For example traffic circulation can be forbidden in certain areas for a number of days.

The French capital suffers from episodes of high ozone pollution each year during the Summer months. The classical symptoms leading up to an episode of pollution are high temperatures (from about 25°C upwards), a lot of sunshine and very little wind. The resulting high level of ozone pollution can have detrimental effects on the population's health. Indeed, ozone pollution being a relatively modern problem, we are still not sure

about the long term effects on the population's health. We hear about Paris a lot mainly because it is the capital. However, other towns in France are coming into the news due to this phenomenon, such as Strasbourg, Montpellier and Lyon. Even fairly modest sized towns are now equipping themselves with pollution monitoring stations. The town of Saint-Etienne is the regional capital of the Loire department in France, it is located about 80 kms to the West of Lyon and counts about 200,000 habitants in the town itself and the suburbs. Our study is based on observed data coming from Saint-Etienne.

Most methods of prediction proposed in the literature are based on statistical approaches, fuzzy logic or neural network based regression. Some laboratories even go to the extent of 3D modelling of atmospheric physics, fluid flow and chemical processes involving partial differential equations and needing a lot of computer power and a lot of time to solve as can be seen in Mounier et al. [2001]. In spite of all these attempts, no method provides really satisfactory results. We insist that we are not criticising the researchers working in this field, we ourselves have tried different approaches but they have not produced the results

so much waited for as can be seen in Pearson et al. [2000], Peton et al. [1998] and Peton [1999]. It is simply a fact that the strong points and weak points of each approach seem to cancel themselves out and we are left with similar results each time.

We believe that, perhaps, the time has come to radically change our approach. The idea is simple, analyse the data collected over the last few years without having any sort of *a priori* model in mind. The fact that a certain number of pollution monitoring agencies have been installed for 5 years or so means that now we are starting to build up large databases covering many different situations. For example, our database from Saint-Etienne started in 1996, meaning that we have a good 5 years of data. A lot of the typical situations leading to pollution episodes in Saint-Etienne are included in the database. A situation we think of as a case and we can therefore apply a case based reasoning type algorithm.

## 2. CASE BASED REASONING AND CELL MAPPING

As explained in the introduction, we have meteorological data available (temperature, humidity, wind velocity, ...) as well as atmospheric pollution data (ozone, carbon monoxide, ...). The data are usually sampled at a rate of at least once every 15 minutes and sometimes even more than that. Not only do we have observed data, we also have forecasted meteorological data, this is an important point for our proposed method. We can use as many or as few variables as we wish, we can use the data raw or filtered, just as we please. The important point is that at round about 16h GMT on day D we need to predict the maximum level of atmospheric ozone for day D+1, the maximum level is usually observed towards 12h GMT but this does not affect our method.

Under the assumption that the natural phenomena producing atmospheric ozone pollution repeat themselves we feel justified in adopting a case based reasoning approach. A case is simply a set of data values and implies the maximum value for the following day. Experience has shown that, for the majority of towns in France and Saint-Etienne being no exception, the following variables produce the best results

- maximum observed level of ozone for day D
- forecast mean wind velocity for day D+1
- forecast maximum temperature for day D+1
- forecast minimum temperature for day D+1
- forecast maximum humidity for day D+1
- forecast minimum humidity for day D+1

Although these variables will be bounded in practice, temperatures at Saint-Etienne do not normally go beyond the interval [-10,40] for example, we must technically consider them to be reals. A case then corresponds to a vector of real values and so we cannot expect to find many days that are identical in this respect, i.e. it is unlikely to find many identical cases. We need therefore to bring some sort of metric into play and this is where we make use of some cell mapping principals.

Basically, a cell corresponds to a neighbourhood in a space and we consider how neighbourhoods are mapped to neighbourhoods rather than considering points being mapped to points as in classical analysis. For a variable  $x$ , we associate to it a value  $z$  when the value of the variable lies in the (half-open) interval  $x \in [ (z-1/2)\delta x , (z+1/2)\delta x )$ , where  $\delta x$  is a parameter determining the granularity of the discretisation. We use the variables listed above and turn all the data into cells in this way, a cell simply corresponding to a vector of integers and representing a case.

Now consider a cell represented by the vector  $u$ , i.e. a vector of integer values. For any other cell,  $v$ , we can calculate its distance from  $u$  via the metric  $\|u-v\|_\infty$  which is just the maximum absolute row sum. We make use of this particular metric because if for example  $\|u-v\|_\infty \leq \rho$  for some positive parameter  $\rho$  then we know that each component of  $v$  differs from the corresponding component of  $u$  by at most  $\rho$ . If we choose  $\rho$  to be an integer then we see how this metric can be used to find neighbouring cells for a given cell. Our logic at this point is of Aristotelian proportions, we assume that for a given cell with a known maximum ozone level at day D+1 a neighbouring cell will produce a similar maximum ozone level at day D+1.

For our preliminary investigations we adopt the following procedure. Assume that on day D we have a cell vector  $u$  corresponding to the observed maximum ozone plus the forecast meteorological values. We fix a value for  $\rho$  and then search in the database for all cell vectors  $v$  satisfying  $\|u-v\|_\infty \leq \rho$ . We then visualise the ozone levels for the days D+1 for all of these vectors  $v$ , usually this is best done with a histogram. Ultimately, we hope that these histograms will supply us with extra information concerning the reliability of the predicted ozone level. We are also looking at the possibility of determining an inclusion mapping  $\phi$  such that  $\phi(u) \subseteq S$ , for all cell vectors  $u$  and where  $S$  is a set of values.

### 3. RESULTS

We applied our proposed method to the database of Saint-Etienne. The raw data run from 01/01/96 to 23/10/01 and we used hourly observations. In reality, we are only interested in the Summer period but for the purposes of this study we used all the data available to us.

One very important problem, one that will be a main target for our future research, is what to do with missing data. Measurement stations can be capricious, sensors can fail and sensors need to be calibrated. The result is a database full of gaping holes. For this preliminary study we applied a very simple filter to the data. If any variable missed more than 6 observations in a 24 hour period we discarded the data for that day. Thus we started with 2122 days of potential data and after discarding the days where there were too many gaps we finished with 774 days of usable data.

We chose the following values for the granularity parameter  $\delta x$ , the data ranges for the whole database (2122 days) are indicated in brackets (wind velocity is used in x-y component form)

- ozone  $\delta x = 10$  , [0 , 260]
- wind velocity  $\delta x = 1$  , [-14 , 10]
- temperature  $\delta x = 2$  , [-10 , 36]
- humidity  $\delta x = 2$  , [15 , 102]

In order to test the method and see what sort of histograms it produced we decided in the first instance to look at all the days amongst the 774 retained when the maximum level of ozone was greater than  $200 \mu\text{g}/\text{m}^3$  (a very high level for Saint-Etienne). There were 4 such days in the retained database. Applying our method to these 4 cases with  $\rho=2$  we produced the following histograms in figures 1 to 4, note that the abscissae in these histograms are all in cell units, the values need to be multiplied by  $\delta x$  to bring them back into  $\mu\text{g}/\text{m}^3$ . One thing that we notice from these histograms is that the columns to the extreme right correspond to the particular days when the ozone value was greater than 200, i.e. the case. All the other values come from days judged similar to the case by our metric based algorithm and, in fact, all of these ozone maxima are greater than  $100 \mu\text{g}/\text{m}^3$ . We believe this result to be promising because  $100 \mu\text{g}/\text{m}^3$  is still a high level of pollution and somewhat justifies our assumption of repeating phenomena.

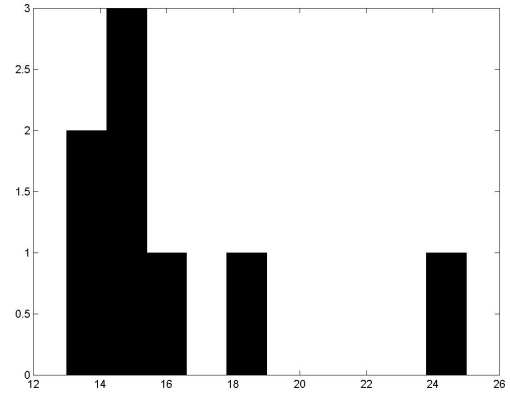


Figure 1 - Histogram of ozone levels case 1

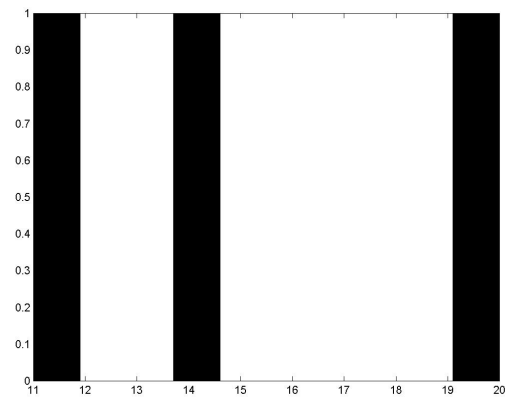


Figure 2 - Histogram of ozone levels case 2

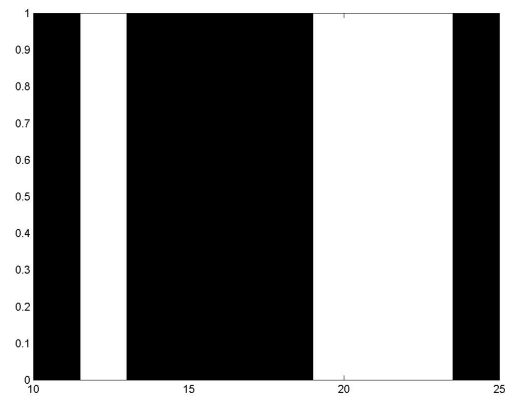


Figure 3 - Histogram of ozone levels case 3

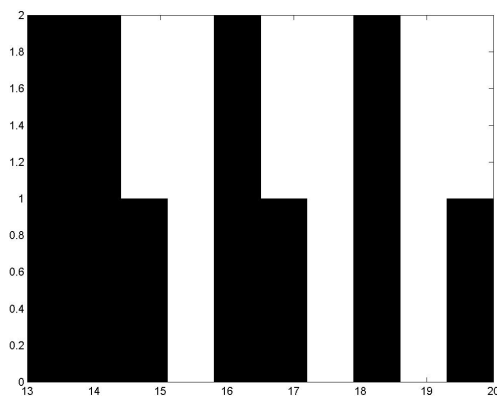


Figure 4 - Histogram of ozone levels case 4

In order to see what happens for another type of day, when the ozone level of day D+1 was less than  $100 \mu\text{g}/\text{m}^3$  we chose another day from the 774 possible and applied the algorithm once more. The result can be seen in figure 5, this time the ozone levels are in general weaker than those in figures 1 to 4, adding a little more support to our assumptions.

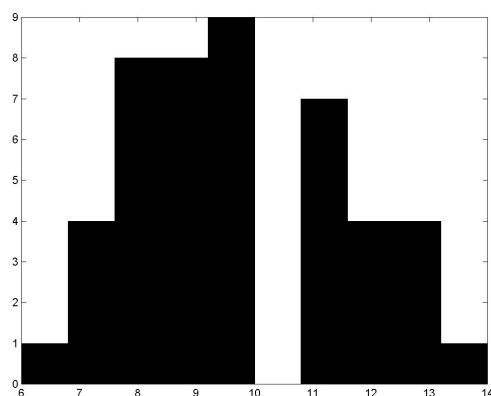


Figure 5 - Histogram of ozone levels case 5

#### 4. CONCLUSION

A new approach to atmospheric ozone pollution prediction has been proposed in this paper. We believe that the approach is promising and we need to carry out more research in order to produce a fully operational software tool.

In fact, what is more promising for our approach is that the new fields of datamining and data reduction are reaching maturity and we are now looking at how we can apply these methods to the pollution problem.

#### REFERENCES

- Mounier G., Couach O., Batton-Hubert M., Clappier O., Photochemical eulerian modelling using multineesting methodology, application to Rhône-Alpes district, paper presented at the 2<sup>nd</sup> International Conference on Air Pollution Modelling & Simulation , INRIA-ENPC, Champs sur Marne, France, 2001.
- Pearson, D.W., Dray, G., Mesbah, M. and Vuillot, B., Ozone, Systèmes Dynamiques et "Shadowing", paper presented at the Journées thématiques "Automatique et Environnement", Nancy, France, 2000.
- Peton, N., Dray, G., Pearson, D.W., Mesbah, M. and Vuillot, B., Modelling and Analysis of Ozone Episodes, paper presented at the International Conference on Air Pollution Modelling and Simulation, Paris, France, 1998.
- Peton, N. ,Méthode de Groupement par Soustraction pour l'Identification de Modèle Flou: Amélioration et Application à la Prévission de la Pollution Atmosphérique, doctoral thesis presented at the University of Montpellier II, 1999.