Helge Rørdam Olesen

June 11, 1999

Detailed account of problems encountered in determining CY's for Kincaid

Overview

I have tried to create an "authorised" data set with quality flagged cross-wind integrated concentration (CY's) for Kincaid, but have encountered several problems. As a result I have only reached a point where I have a *provisional* data set with CY values. An important lesson learned from the work is that the definition of arcs in the KININPUT.BAK data set set ought to be re-examined. Another important observation from my work is that the crosswind integrated concentrations computed by John Irwin's NEWFIT programme overestimates the true CY's.

A possible point of departure - Steve's data

In Steve Hanna's Kincaid data set which I received for the Manno workshop in 1992/93 there was a file (distributed for the Manno workshop as ARC_KIN.DAT, size 1.2 MB) which contains all concentrations from the Kincaid set, arc by arc, as well as some information relevant to the crosswind integrated concentrations along the arcs (but not the computed CY's themselves).

Example:

```
80  4 20 14 AZ:   85   87   88   90   92   95   96   99  102  105  ...
R =  3.0 KM;OBS:  46.    0    0   2. 127.   1. 125.   7.  73. 104. ...
CY CUM FRACTION: .000 .031 .031 .032 .119 .249 .291 .425 .505 .684 ...
```

Among Steve's data, there is another data set which actually contains arc-wise CY's. The problem with Steve's CY data is that they are computed by an automatic procedure, disregarding their quality (no proper handling of incomplete arcs, gaps in arcs etc.). They were therefore omitted from the versions of the Model Validation Kit which were distributed after Manno.

I do not have access to the arc definitions used by Steve (the above mentioned file ARC_KIN.DAT does not contain the precise position of monitors, only their azimuth and an approximate distance). This information could probably be derived by re-running some of Steve's old data generating programmes, but so far I have not pursued this thread because there was another alternative.

Another point of departure - John's data

I chose a different route in determining the CY's, based on John Irwin's KININPUT.BAK data set. However, I used Steve's CY's values for comparison purposes (and this eventually led me on the track of problems with John's arc definitions).

I assumed that the advantages of using John's data would be compatibility with his set of procedures for ASTM computations, and that I would be able to largely re-use his Fortran programmes.

So, based on John's NEWFIT.FOR programme I created a programme (NF_AREA.FOR) which would give me the CY's for each of the arcs in KININPUT.BAK.

During this work I discovered a danger in applying NEWFIT.FOR for computing CY's. The programme is not very well documented, so the danger may pass unnoticed. The problem is that CY as computed in the BESTFIT routine is ***not*** the true trapezoidal integral of crosswind concentrations. Let us call the BESTFIT CY for CY_JI and a true trapezoidal integral for CY_TRUE. Actually, on the average for all (approximately 650) arcs when integration was successful, CY_JI = 1.4 * CY_TRUE, so John's CY values represent a considerable overestimate. This trap has also implications for all values based on Cy (such as the fitted CMAX). However, it is not clear to me whether this is of consequence for any of John's main results.

When I made computations of CY_TRUE, I replaced the routine AREA by a routine of mine called AREATRUE. The section below is taken from my comments in the AREATRUE Fortran code and explains briefly what the problem is with the AREA routine.

```
AREATRUE is based on John Irwin's AREA routine, but is simplified.
AREATRUE performs a TRUE trapezoidal integration as opposed to AREA.

AREATRUE does not require much explanation, whereas AREA does
-  so here are a few comments on AREA.

AREA assumes that the profile to be integrated is convoluted,
and that it has been formed by taking absolute values of Y
(note that Y is the INDEPENDENT variable, i.e. the distance from
the centre of the profile).
AREA further adds a 'front part' and a 'tail' to the integral,
and finally multiplies the integral by 2.
Both the folding procedure and the addition of head and tail
lead to a larger integral than a true trapezoidal integration.

AREATRUE makes no assumptions, but computes a plain integral.
```

My procedure for CY quality control

In the following I outline the way in which I created a data set with CY values. As it appears, I am not satisfied with the result and will not recommend that the resulting data set is used as an "authoritative" one. The steps I took were the following:

I derived CY data from KININPUT.DAT by running my program NF_AREA.FOR. NF_AREA.FOR is based on NEWFIT.FOR and applies the same criteria for determining whether an arc is suitable for integration (a minimum number of monitors etc.). It results in both a true CY and a CY_JI. The immediate result is a large file (around 1 MB) with 15000 observations, one observation for each monitor where integration has been successful. There were 651 arcs for which this was the case.

These CY data were merged with data from the Model Validation Kit. In the Model Validation Kit, the file SF6_KIN.DAT lists data for 1284 arcs, each with an observed maximum. In the data of the kit, a quality indicator QUAL (values 0-3) is been assigned to each observed arcwise maximum, indicating its reliability. Approximately half of the arcwise maxima (586) are of good quality (2 or 3). I would like to retain this information in the resulting data set. Therefore, at a this point in the process, I ran a few programmes (written in the SAS language) in order to merge the Model Validation Kit data with the CY data derived from KININPUT.DAT. (I chose SAS for this purpose because SAS is suitable for dataset merging though it is a bit heavy to work with).

I further merged the data with Steve's original data for CY. The merging was performed such that, e.g., data for John's 5-km-arc for a certain day and hour were merged with Steve's and the Model Validation Kit's data for the same 5-km-arc. Thus, I neglected the fact that the monitors in the arcs are not necessarily precisely the same.

As a matter of interest I also merged information on each arc's regime according to the definition given in John's KKZIOL2.TXT file (this information may be instructive but does not belong in an authoritative data set).

I made plots of each of the 651 arcs and inspected them manually in order to assign a quality indicator to the CY value (CY as derived by trapezoidal integration). I assigned each arc a quality indicator from 0 to 3 (QCY) as indicated in the text below:

Contents of the file CY_QUAL.TXT

Quality criteria - explanation of the indicator variable QCY

Cross-wind integrated concentrations (CY) have been computed using a simple trapozoidal integration along "John Irwin's arcs", i.e. arcs as they have been defined in the NEWFIT programme. A quality indicator is assigned to such CY values.

One of the main criteria for judging whether the CY value should be considered reliable is whether the arc is "complete", i.e. whether zero values are included at the edges of the arc.

The variable QCY can take on values from 0 to 3. For model evaluation, users should consider only CY with a quality indicator of 3, or possibly both 2 and 3. More specifically, the values of QCY have the following significance:

0: A computed cross-wind integrated concentration is missing or the value should clearly be disregarded. This includes 3 cases with outliers. See the file OUTLINEW.TXT for details.

1: The estimated CY value is not reliable. Reasons which may lead to this judgement are:

- The arc is obviously not complete
- There are severe gaps in the monitoring arc
- Comparison with other arcs during the same experiment (their magnitide or the plume direction) gives strong reasons to suspect that the arc is incomplete.

2: The arc seems complete or almost complete, but there are reasons for caution. If you use data of quality level 2 for model validation, don't give too much weight to your conclusions as there are probably many misleading estimates among these values. Quality 3 data are to be preferred.

Reasons for assigning a quality indicator of 2 rather than 3 include:

- The coverage of monitors is not dense (note: this can imply both that the CY as deduced from observations is anOVERESTIMATION of the real CY as the opposite).
- The arc is only ALMOST complete.
- Although the concentration values go to zero at the edges of the arc, the structure of the arc is such that there may well be non-zero concentrations beyond the edges

.

3: The arc seems complete and can be used for model evaluation (zero values or close-to-zero values are present at the ends of the arc, and there is a reasonable dense coverage of the arc).

Note: It is not required that the cross-section is well-behaved and has only one significant peak. It is not required that the variation of CY with downwind distance is as should be expected (that the CY rises to a level where it stays)

The manual inspection was almost entirely based on inspection of plots like Fig 1 showing crosswind profiles. This work could be relatively quickly done (less than one day of work for the 651 plots). The drawback of the simple method with inspection of this type of plots is that it may pass unnoticed if an arc is improperly defined or if it is completely misplaced in respect to the plume.

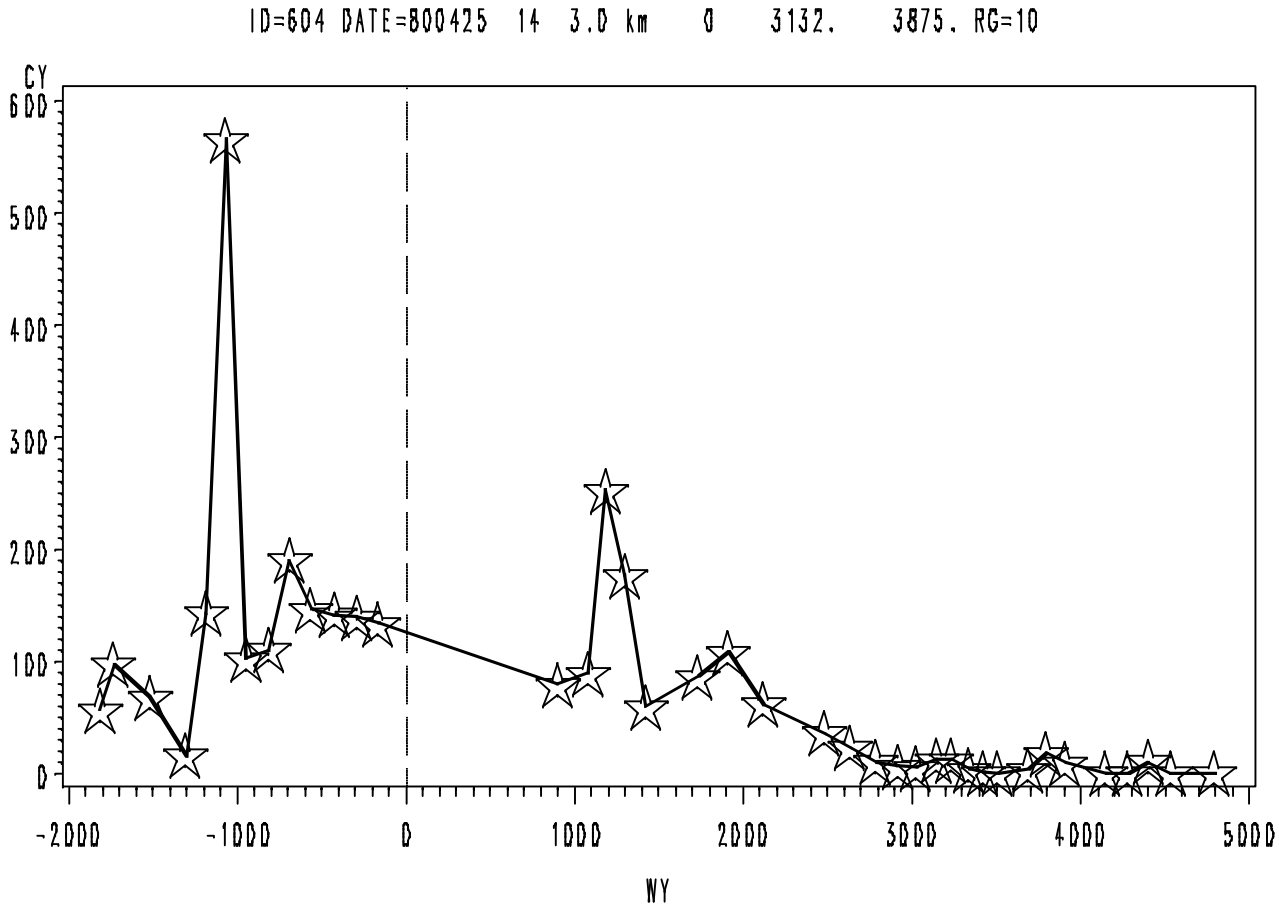

Fig. 1 April 25, 1980, 14 hours. When inspecting crosswind profiles this arc was assigned a quality indicator of 2 (in respect to cross-wind integrated concentration, i.e. QCY=2). This value was chosen because the arc is only *almost* complete (it does not go to zero). Further, the arc of monitors is not dense near the centre of the arc. At a later step in the quality control procedure I realized that the definition of this arc was problematic (see the example in a later section).

As a further quality control, I compared Steve Hanna's CY values (CY_HAN) with the new ones. There were discrepancies which led me to consider some cases in more detail, and eventually to ***the conclusion that the definition of arcs ought to be re-examined***. Only then can we have a trustworthy set of data. I will show some examples of the problems in the next section.

The provisional outcome of my efforts is a data set where some misleading data are included. I have, however, marked the least trustworthy data by special values of a quality indicator.

I have added 100 to the quality indicator for those arcs where there is a large difference between Steve's CY values and the KININPUT.BAK based CY values. I have defined a "large difference" as 25%, implying that 113 of 651 arcs have been flagged with a quality indicator greater than 100. If I further restrict my interest to data with a quality indicator of 3 (QCY=3), 325 arcs remain. My provisional data set contains observations flagged with a quality indicator QCY, with possible values of 1, 2, 3, 100, 101, 102 and 103.

Problems with the arcs in John's data

1. Examples of arcs with an inappropriate definition

In this section two examples are shown where arcs are inappropriately defined.

First example

The first example is for May 4 1980, 10 hours, where some points have been assigned to a wrong arc.

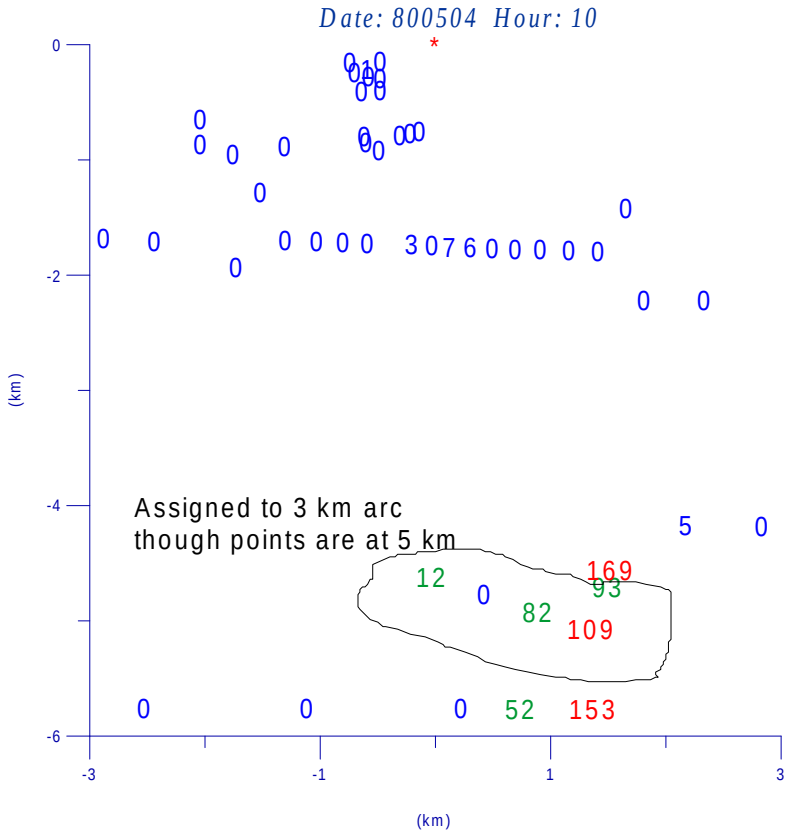

Fig. 2 The points within the line are assigned to a 3 km arc while they ought to be included in the 5 km arc like the values of 169 and 153. Thus, misleading information on the 3 km arc is present in the data set (and insufficient information on the 5 km arc). This misleading information is used in all analyses using the ASTM methodology. Concerning cross-wind integrated concentrations, the particular example of the 3 km arc has been caught by the quality control where I compare the computed integral with Hanna's old value and find the difference unacceptably large.

Second example

A second example is from April 25, 1980 at 14 hours. A closer examination reveals that the crosswind profile shown in Fig. 1 is not at all representative for concentrations at the 3 km distance. The layout of monitors for April 25, 14 hours is shown in Fig. 3. In fact there are no monitors at 3 km distance close to the plume centre line. Most of the monitors assigned to the 3 km arc are at 5 km distance from the source.

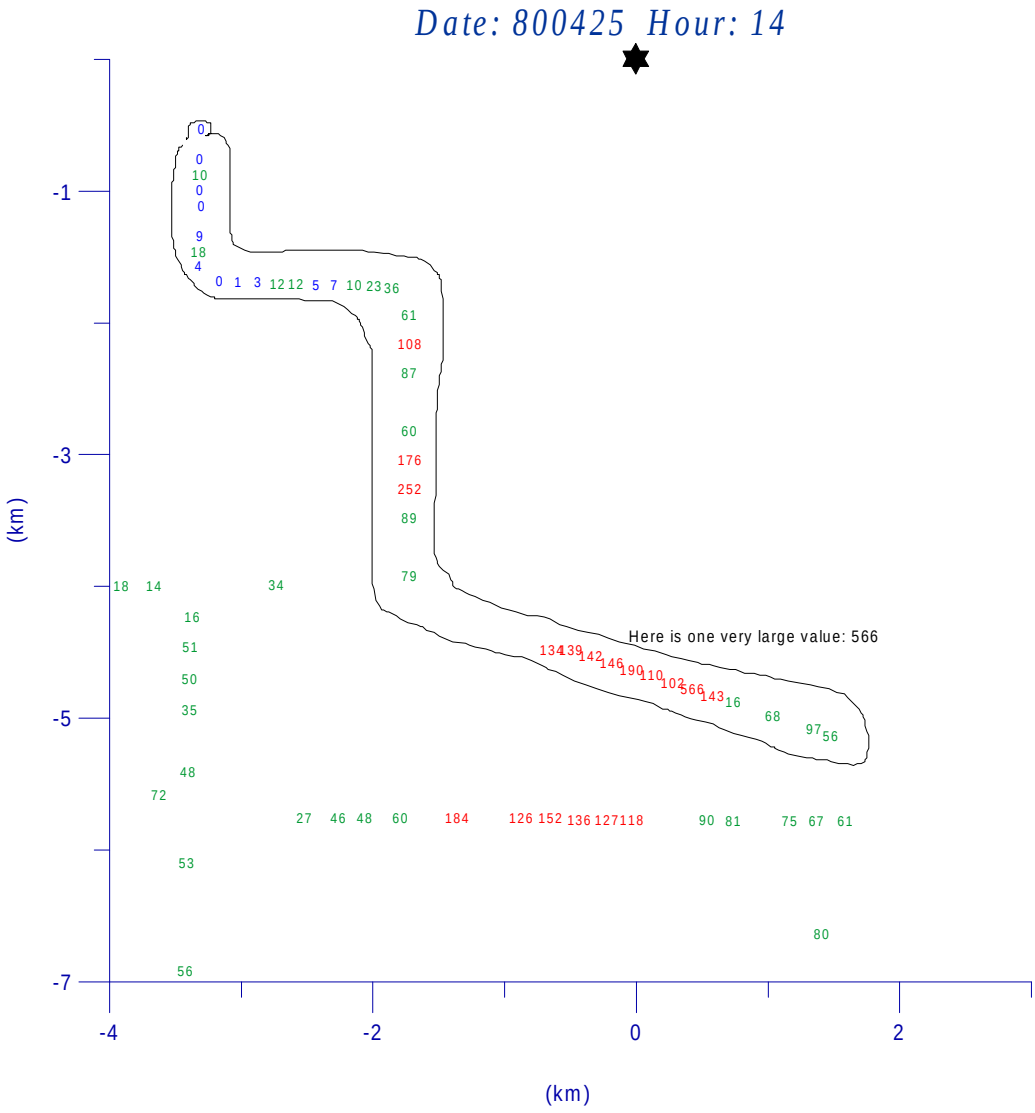

Fig. 3 The points enclosed in the line have been assigned an arc distance of 3 km though most of them should rather be included in the 5 km arc.

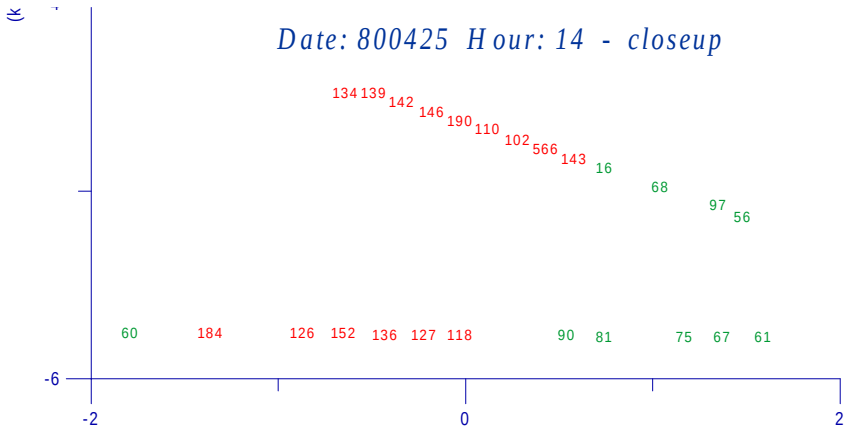

Fig. 4 Close-up of a limited area from Fig. 2.

2. Arcs with "distance" indicated as 99

In the KININPUT.BAK file for Kincaid there are arcs with distance indicated as "99".

I do not know the criteria for assigning this value to observed concentrations, but it does NOT mean that the distance to the monitor is 99 km - rather that it is not a member of the regular monitoring arcs.

The value 99 for arc distance occurs for May 22, 1981 and some of the following days in May/June 1981 (and also once in 1980, namely at July 11, 13 hours).

Examples of monitors with this value is monitor at a distance of 6.07 km (which is a bit apart from the 7 km arc), and a monitor at a distance of 14.82 km (which is at an almost identical position as another monitor; maybe they are too close to one another for some programme to work, so one of them has been discarded?).

In my treatment of data, I have assigned a quality indicator QCY of 0 to arcs with a distance of 99.