Problems with Calculations of Cosmic Ray Shower Direction

Michael McEllin

1 Summary

One of the design intentions of the HiSPARC cosmic ray detector network was the use of GPS-timestamped data from multiple HiSPARC stations to learn more about cosmic ray 'Extended Air Showers'. Examples, software tools and reports on the HiSPARC website encourage students to solve, in particular, for the arrival directions of EASs. The validity of these directional solutions rely heavily on the accuracy with which EAS event times are measured, and systematic errors in the timing circuits would be especially misleading since it would strongly bias the direction solutions.

I present some preliminary evidence that all may not be well with this timing information. There could well be some moderately large systematic time errors that have a serious impact on the direction solutions. This needs further investigation, which is possible by looking more closely in to the information on the HiSPARC database.

2 Background

Cosmic ray 'Extended Air Showers' (EAS) trigger coincidence events between the detector plates at individual HiSPARC stations. Events are timestamped with the arrival time of the shower, using a time taken from GPS signals, which are available at all HiSPARC stations, and the information is passed to the central HiSPARC database, from where it can be downloaded by any interested partly (using software supplied by the HiSPARC project). I will refer to these in-station coincidence simply as *events* (meaning EAS events) in the rest of this document.

Most HiSPARC stations have two detector scintillator plates, which if triggered sufficiently close together provide confidence that the station is detecting a real cosmic ray event. (This means that cosmic ray particles are

arriving in a bunch too close together to be considered to have independent origins.) Two detector plates, however, provide little information about the direction from which EAS arrived[1]. Some of the HiSPARC stations in the Netherlands, particularly those located in Nijmegen Science Park, have four detectors spaced in such a way that the relative timing of the trigger does give directional information. (Fokkema 2012) This information has been validated as relatively accurate by comparing data from a four-detector-plate station with that from a professional detector installation (Fokkema 2012).

Sometimes a sufficiently large EAS will trigger events at more than one HiSPARC station. This is also known as a multi-station coincidence. One can recognise such coincidences by searching through the database of in-station coincidences and, using the GPS timestamps associated with each event, searching out those events that occur sufficiently close together in time. Since the shower front moves at very nearly the speed of light (and is fairly thin—rather like a circular plate) if the time difference between events registered at two stations is less than the light-travel time along the line between the stations then there is the possibility that the stations are registering the same EAS. Such coincidences might occur by chance, of course, but knowing the average rate at which events are detected we can calculate that this is fairly unlikely. For the rest of this document I will simply refer to these multi-station EAS detections as *coincidences*.

Software supplied by the HiSPARC project can be used to interrogate events in the HiSPARC database and identify such coincidences. As it happens, for the 'Science Park' cluster in Amsterdam, this analysis has already been performed and the results recorded in the HiSPARC database in a 'coincidences' table. This means that we can use the Python 'Sapphire' module, (supplied by the HiSPARC project) to directly download 'summary' data for coincidences for the eleven stations in this cluster. This provides us, for each coincidence, with the number and identity of the HiSPARC stations calculated as being in coincidence (so we can look up their exact positions as recorded by GPS devices) the times of arrival, as determined by GPS, at each of these stations, and a reference to other data tables, providing more information about the triggered events at each HiSPARC station (should we

[1]We would get the same time difference between the detector triggers with an arrival track anywhere on the surface of a cone with its axis aligned with the line joining the two plates. It is, however, somewhat more likely than not that the detected EAS comes from directions closer to the zenith direction, for two reasons: the detectors, being flat horizontal plates, are most sensitive in this direction and showers travelling at a large angle to the zenith have to travers a greater depth of atmospheric material and are more likely to be absorbed.

need it).

If we have at least three stations involved in a multi-station event, then in principle we can use the arrival times at the three stations to determine the arrival direction of the shower. A number of examples on the HiSPARC website illustrate this as a useful exercise for students, and the Sapphire Python library contains procedures to help with the processing. I think, however, that the evidence presented in this note suggests that there is little point in pursuing such experimental analysis until the validity of the raw data has been further considered.

Simple trigonometry shows that if the shower is approaching at an angle θ to the line joining the stations, then the difference in arrival times (in seconds) at two nearby stations, separated by distance D, will just be:

$$dT = \frac{D}{c}.\cos(\theta) \tag{1}$$

where 'c' is the velocity of light. In future I follow a common convention amongst high-energy physicists and choose to measure distance in terms of the time it takes light to travel between the two points. In fact, given that time differences are reported by HiSPARC in nano-seconds (that is 10^{-9} seconds) we might as well choose one nano-light-second as our unit of distance and just get:

$$dT = D.cos(\theta) \tag{2}$$

Clearly, dT has to be less than D in order for θ to be a valid angle. In fact the algorithm that populates HiSPARC coincidence database table filters out any data that do not satisfy this condition—we simply never see them in the list of multi-station coincidences. With three stations we should be able to calculate two angles against two inter-station baselines, and it is easy to see that there are only two directions in space which can make both angles valid. We resolve the ambiguity because one of these points into the ground and the other, the one we want. points to the sky. This calculation relies, of course, on an accurate measurements of the time delays. Any error in measuring the relative times will produce an error in the calculated EAS direction.

There are several reasons why such errors might occur:

- GPS signals are subject to delays as they travel through the ionosphere, with the amount of delay dependent on the electron density along that particular track. Although average corrections can be applied there will be some variation, which may change from day to day. The amount of time delay also depends on the frequency of the GPS

transmissions, so sophisticated and expensive GPS receivers, such as those used by the military, use GPS signals on a number of different frequencies to calculate a very accurate correction. We do not have this capability.

The sources of error in GPS times are complex: some are random, some systematic depending on time of day, and some systematic in the GPS receiver. The bottom line is that we cannot rely on them to better than about 40 nano-seconds (and though they can occasionally be better, sometimes they are worse that this). However, we might expect that for HiSPARC stations that are relatively close together the systematic errors in propagation time are likely to be very similar for each stations. If this is the case then the relative difference in arrival times of an EAS at each station is unaffected.

In practice most coincidences are registered for detector stations separated by only a few hundred meters at most. It seems very likely that differences atmospheric propagation time is unlikely to be a significant source of systematic errors.

- The positions of the detectors (measured with the GPS system) may be incorrect. Civilian GPS receives generally produce results accurate to within a few meters on a single observation, and the HiSPARC locations have been calculated by averaging position measurements over an extended time to get better accuracy. This probably reduces any systematic time errors due to position accuracies to the nano-second level. (Light takes 3 nano-seconds to travel one meter.)

- The particles in the EAS shower front are not exactly in one plane, so there is some variation in arrival times. (The shower plane has a certain thickness, and it is also slightly curved with a radius extending back to the point where the original high-energy cosmic ray encountered the first air molecule.) These errors, however, should be small.

- The scintillation detector has to register a light pulse that rises from zero to a peak and then decays. The arrival time has to be defined at a trigger point during the rising phase of the pulse. Any difference in trigger levels, or the shape of the pulse generated will lead to timing errors.

- The electronics that processes the signals from the detectors will inevitably introduce delays. Electronic components are not absolutely

identical so the amount of delay introduced before a GPS time signal is registered may differ between installations.

For the last two potential sources of error, in particular, we have no method of quantifying the possible magnitude, and there is no information on the HiSPARC website.

3 Predictions and Tests

Extended air shower arrival directions, in general, are known not to have any strong association with particular directions in the sky—though at the very highest EAS energies, where interstellar magnetic field have little effect on their propagation directions, large professional cosmic ray detectors may have established a connection between their arrival directions and positions of active galactic nuclei (AGNs).

At the typical energy of EAS events detected by the HiSPARC network, however, it seems likely that the interstellar magnetic field is likely to randomise the arrival directions to a considerable extent. We would therefore expect to see a plot of arrival directions to appear evenly distributed in azimuth.

We *do* expect to see correlation of frequency with zenith angle, because the HiSPARC detectors are flat horizontal plates, and their projected area in the direction of shower arrival falls off with a $cos(\theta)$ variation (where θ is the angle of the shower track with the direction to the zenith, i.e. the point directly overhead). In addition those showers incoming at bigger zenith angles travel through more atmosphere and are therefore more likely to be absorbed before reaching the detector. The greatest frequency of events will therefore be from directly overhead, falling off somewhat *more* steeply than $cos(\theta)$. We would need to examine departures from the predicted variation of event rate to detect non-randomness.

Nevertheless, after plotting the positions of all 3-station coincidences from the Science Park cluster in Amsterdam between January and October 2017 in an Altitude-Azimuth plot some anomalies are clear. (See Figure 1, in which every dot represents a coincidence event.) The centre of the plot (where the red cross-wires meet) represents the zenith, and the horizon is a circle with a diameter almost as large as the frame. As explained earlier, the clustering of data towards the zenith is much as expected, because the HiSPARC flat-plate scintillation detectors are most sensitive in this direction. However, there appears to be a distinct off-centre bias and some odd radial spurs in which data-points appears to be largely absent.

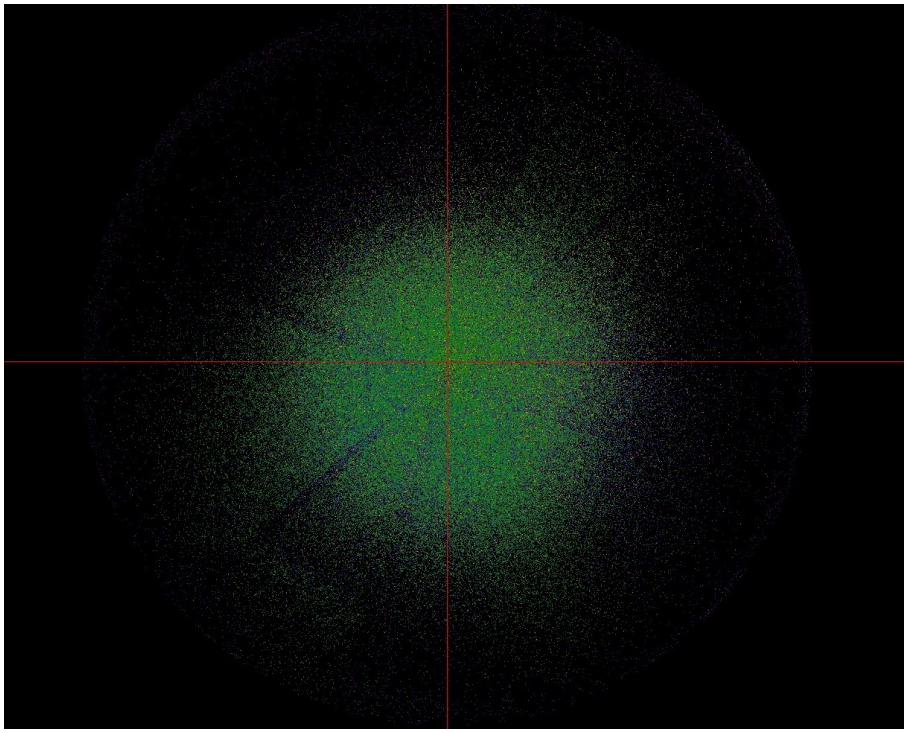

Figure 1: An Altitude-Azimuth Plot of CR Arrival Directions.

Anyone familiar with software construction would immediately suspect an error in the direction solution algorithms—in fact I produced this plot because it was likely to reveal any systematic errors in the calculation. Until we are certain this is not the case that should remain under suspicion, but I think that at least some of the evidence now points in a different direction.

How do we investigate this?

Figure 1 aggregated all the direction solution data from *all* 3-station coincidences detected by the Science Park cluster. We might get more clues about errors in the analysis if we specifically focus on just three specific stations at a time and plot all the coincidences from just these stations. If there is odd behaviour in the direction solution algorithm it may well show up more clearly in unexpected patterns in a more limited context.

Figures 2, 3 and 4 show three examples of this type.

The first appears to show a well distributed set of data—but with a distinct off-centre bias to the south west. (Note the the three green dots

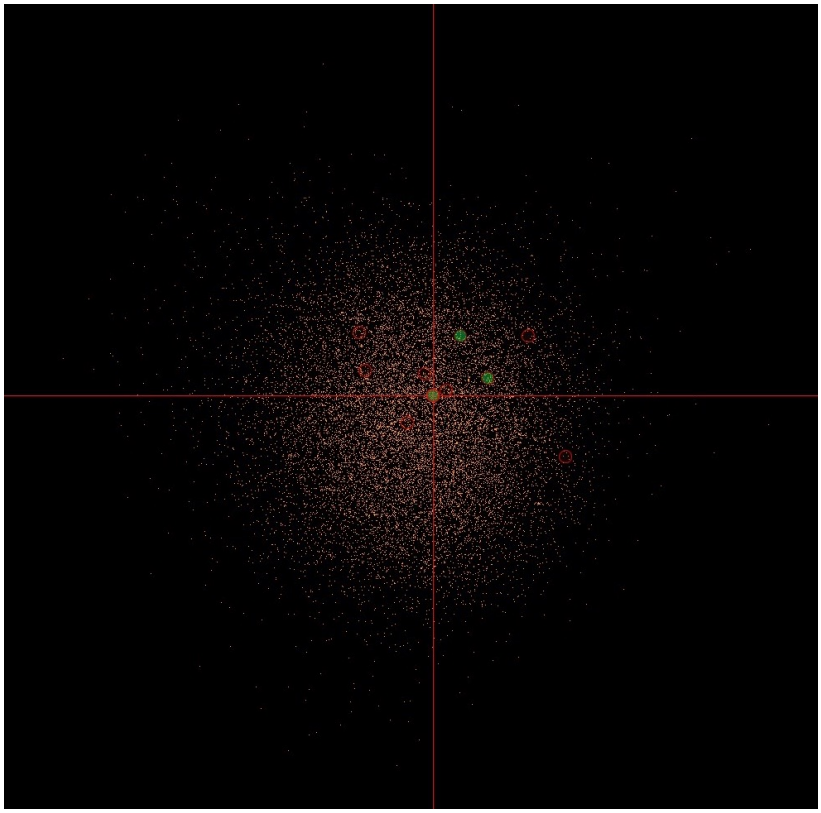

Figure 2: An Altitude-Azimuth Plot of CR Arrival Directions - Station Configuration A.

show the relative positions of the three Science Park stations used to collect the data. The hollow red circles represent the relative positions of the other Science Park stations.)

This is the type of pattern that might occur, if, for example the station at the centre of the crosshairs was represented as being slighter faster in registering an event that the other two stations. There are two hypotheses: either the software processing is introducing such a bias, or perhaps the station hardware is actually introducing a time bias at the point when the event was detected.

Figure 3 is more difficult to explain. Here we see a clear absence of any results in one entire sector of the sky. One hypothesis here is that an even larger systematic timing bias in one of the stations means that EAS from

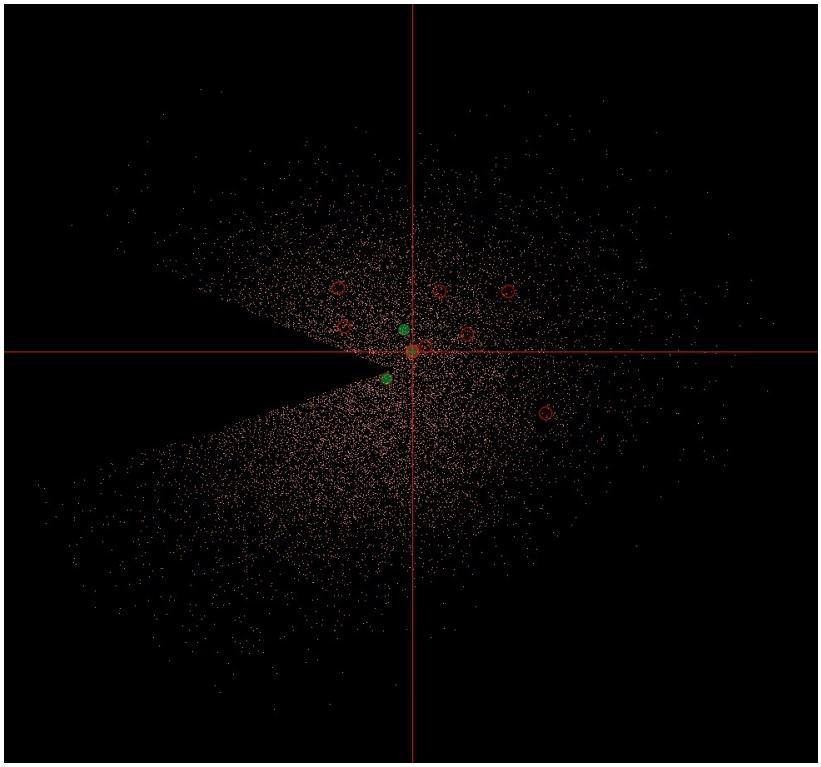

Figure 3: An Altitude-Azimuth Plot of CR Arrival Directions - Station Configuration B.

the North West are rejected as physically infeasible, because the systematic time error is larger than the distance between two of the closely separated stations near the plot centre.

Figure 4 also shows a distribution which missing data—though this time it is in a larger sector. Again, I suspect that data from this particular direction is being rejected as physically infeasible by the filtering algorithms.

At this point I cannot conceive of any class of software error that produces this type of pattern in the results other than systematic offsets in the timestamp values used for particular stations. In spite of a great deal of close inspection I have not been able to identify any such error as yet since would it requires a type of mistake that picks out particular stations for different treatments: that ought to have a very specific signature in the code. (I can always be proved, of course, to be suffering from a lack of cleverness

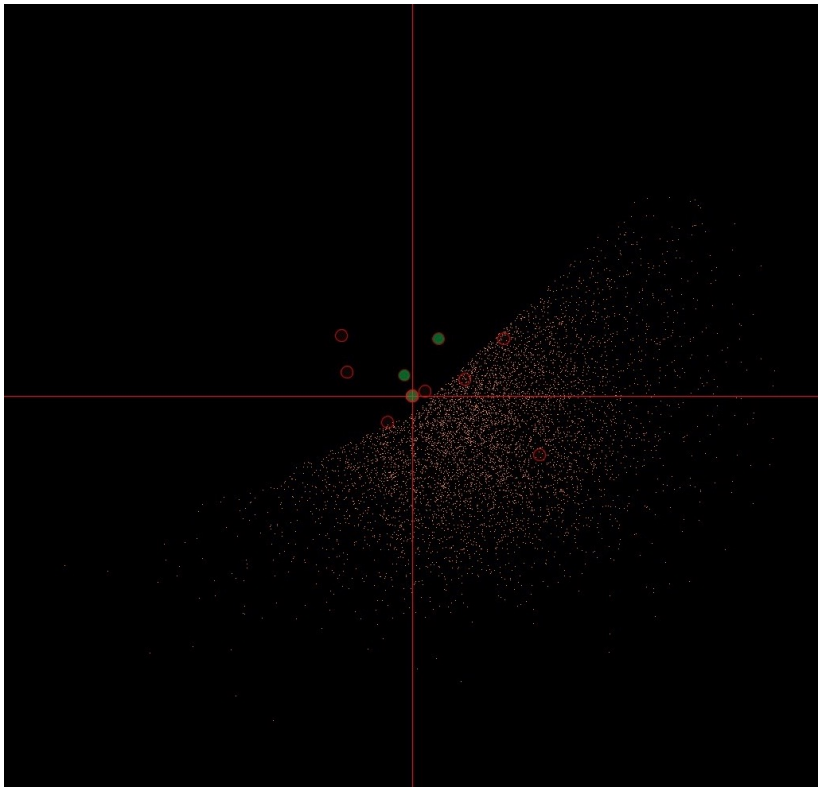

Figure 4: An Altitude-Azimuth Plot of CR Arrival Directions - Station Configuration C.

and imagination, but when you have a lot of experience of tracking down data and software errors, you tend to investigate errors in their order of likelihood. A serious coding error is for the moment demoted to a less likely position.)

My primary suspect is now a hypothesis of systematic timing errors in the HiSPARC station hardware that vary from station to station. (That is, each station consistently suffers from the same error, which is different to that at other stations.) We therefore need to think about techniques to confirm that the raw data shows the type of statistical patterns we might expect from a true random distribution.

Note that the Science Park cluster is quite concentrated—stations are only at most a few hundred meters apart—some separated by only 100 nanoseconds or so in light travel time, so a small time error—perhaps 20-

40ns—may make a big different in the angle calculation. (While in principle we can look for inter-station coincidences for more widely separated stations, in practice we would expect to see many fewer positive results because EASs have a limited horizontal extent, generally no more than a few hundred meters. Hence, the bigger the separation of stations, the lower the rate of multi-station coincidence detection. This *is* a pattern we see in the data.)

The closest we can get to the raw data is an examination of the distribution of differences in arrival times at selected pairs of stations, which we can extract directly from the HiSPARC database. (We will assume that the data has been correctly recorded and correctly extracted.) We can make two testable predictions about this data that are easy to check. They are based on the hypothesis that there is no overall genuine bias in EAS arrival directions with regard to azimuth, that is, that they are completely random. (As explained above with *do* expect some concentration of events towards the zenith, but as long as the detector locations do not greatly differ in height this should not influence the logic below.)

From equation 2 on page 3 of this document: we know that the cosine of the angle between the arrival direction and the two-station baseline is just the ratio of time delay over distance. We are claiming that the arrival directions are randomly distributed, so the ratio dT/D should also be drawn from a certain random distribution which in fact should have the *same* shape for any two stations we choose. We can also deduce that the distribution ought to be symmetrical about zero. You may need to think about that, but we can prove it mathematically.

Note that I do NOT mean that any value of dT/D between -1 and +1 is equally likely: there will be a tendency for the most likely values to be clustered around 0. We can calculate an approximation the shape of this probability distribution by ignoring the absorption of low-angle tracks—it is A-Level maths and I do not intend to reproduce it here—though we can see fairly intuitively what the answer has to look like. There are fewer bits of the sky at small angles to the two-station baseline than at large angles, and in fact the sky area goes to zero at zero angle to the baseline so the probability of events at this angle must drop to zero. We might expect that this would apply a $cos(\theta)$ variation, where θ is the angle from a plane perpendicular to the inter-station baseline. The scintillation plates also show a bigger cross-section at small angles to the zenith, so we might expect a multiplication by another $cos(\theta)$ variation. The probability of getting a particular θ is therefore going to be related to something narrower than $cos^2(\theta)$ once we take account of greater absorption for tracks away from the zenith.

The argument above is not rigorous but there is not much room for it to

be significantly wrong. In fact, I have done the 2D trigonometric integrals and it works out as expected! You can often get a good feel for how the answer to a problem should turn out before doing the rigorous calculation, and is a useful thing to do because it often helps you to avoid mathematical slips. It also helps you to really understand the nature of the problem and where it is most sensitive to assumptions.

Hence, it is easy to see that

- The mean value of DT for any station pair should be close to zero (within expected statistical variations). That is it is equally likely that EASs come from one direction along the baseline as for the opposite direction.

- The shape of the distribution of the ratio dT/D for any two stations should be very similar (within expected statistical variation) and that this distribution should be a normalised curve varying $\propto cos^2(\theta)$ or most likely somewhat narrower. (The narrowing would arise because tracks at large zenith angles have a higher tendency to be absorbed in the atmosphere. We have *not* take account of this in the calculation because we have not yet done the experiments to quantify the effect, and it is not easy to predict mathematically.)

I need to explain 'expected statistical variations'. If we have a random variable (such as the ratio dT/D) then by selecting an infinite number of samples from this distribution we would exactly work out the mean value. If, however, we take a finite number of samples (say observations of dT/D) by random chance it is quite possible that we would select a group of values that is slightly biased in one direction or another (and next time we selected a random group it would be biased differently).

Knowing the form of the original statistical distribution we can, in fact, predict just how much variation in the mean we are likely to see when we take finite sized samples. It usually works out to be something close to $\sigma/\sqrt(n)$, where σ is the 'standard deviation' [2], and n is the number of samples in our group. Hence, the larger the value of n the smaller the likely variation in the mean we might calculate (that is, it tends to zero as n goes to infinity).

Note that this probing of the data is *not* dependent on any sophisticated process such as solving for arrival directions. The steps undertaken involve only:

[2]The 'standard deviation' is a measure of the width of a distribution. I will not define it here: you can find it online. Note that tools like Excel provide the calculation of this parameter as a standard option.

- Downloading 3-station coincidences for the Science Park detectors.
- For each coincidence, extract the individual dTs for all station pairs.
- Accumulate histograms of dT for each unique station pair. Note that some station pairs will have appeared in more than one 3-station configuration. For example, we might have S510, S511 and S502 and S510, S505 and S502 where S510-S502 is a shared baseline.
- Plot the histograms of dT/D.

It appears to me that there is clearly an issue with the raw data, regardless of whether I have made any subsequent mistakes in deriving arrival directions.

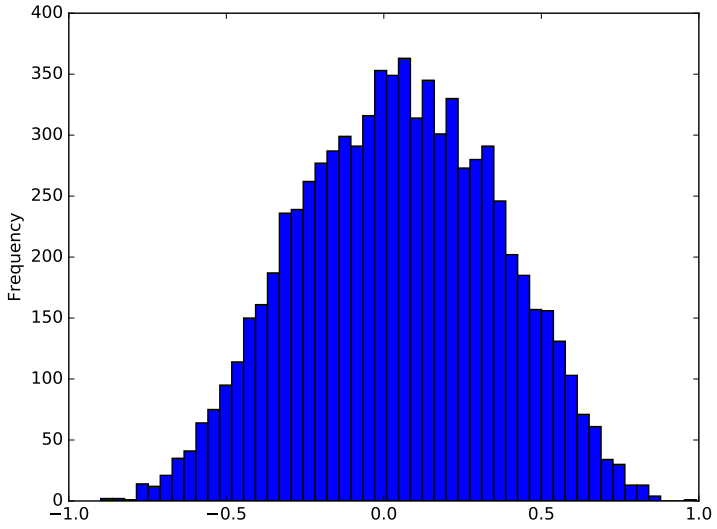

Figure 5: dt/D Histogram for Stations S510 and S505.

Figure 5 shows an example of the distribution of dT/D values where the bottom axis is the ratio dT/D and the vertical measure is the number of examples in boxes each box. Figures 6, 7 and 8 show further examples. The shape shown in Figure 5 is in fact more-or-less as expected apart from a slight bias to the right (i.e. positive dT/D). The bias is small, but given the size of the dataset, I suspect that it is significant (that is, it is unlikely to have occurred due to taking a random sample of this size from an unbiassed

distribution). One would need, however, to positively confirm this with a proper statistical test, and also that the shape really does corresponds to something like the expected profile.

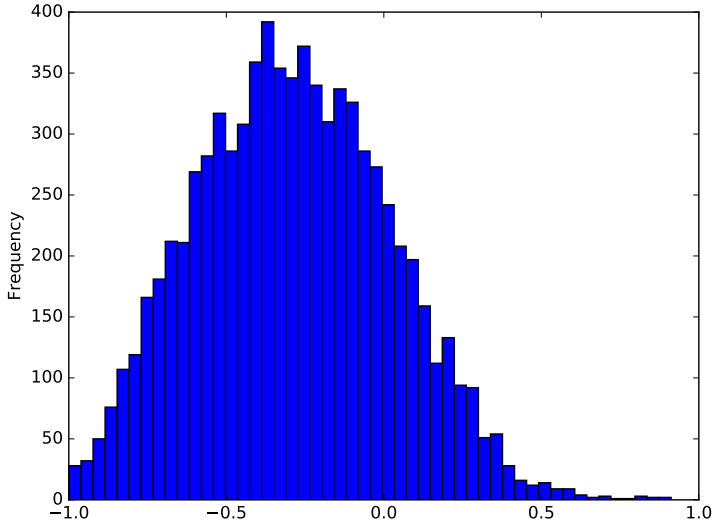

Figure 6: dt/D Histogram for Stations S511 and S505.

Figure 6 has the expected shape, but is very definitely biassed to the left and values with $dT > D$ have clearly been filtered out by the HiSPARC software.

Figure 7 in contrast is perhaps not symmetric, as it should be. Figure 8 is definitely asymmetric as well has having a right bias. (The are statistical tests that look at asymmetries which we could apply to demonstrate this rigorously.)

Although I have not yet eliminated the possibility of a processing error in my own direction solution program—I have been constructing software for far too long to still have any illusions about my own infallibility—the patterns shown here (and in other histograms that I have not included) seem to me to be more likely explained by timing offsets in the HiSPARC detector stations. (In the case where the histograms are clearly non-symmetric it may well be that the time-offset changed during the period of logging—perhaps by alteration of the photomultiplier tube voltates— so that we are in effect

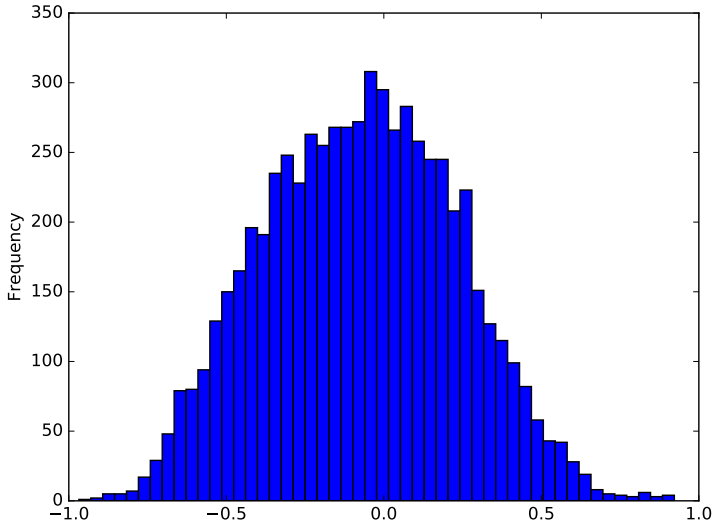

Figure 7: dt/D Histogram for Stations S511 and S502.

looking at two superimposed histograms with different biasses.)

There are several questions that we need to address before we can consider using the HiSPARC data for direction solutions:

- Can the data we see be explained by assigning fixed time offsets to each of the stations on the Science Park? I have not closed the loop and reconciled the expectations that would arise from the biassed histograms with the effect on the distribution of direction solutions. This needs to be looked at.

- Does the truncation of data at $|dt/D| > 1$ help to explain the odd patterns in the direction-solution plots?

- Is there any way to introduce corrections so we can do multi-detector solutions for EAS arrival directions? (I wonder if some of the timing offsets may be dependent on photomultiplier tube voltage settings that have changed over time, so we may not get far with this hope.)

- What happens if we generate a fictitious unbiassed random set of EASs, and process it through the direction-solution software? Will

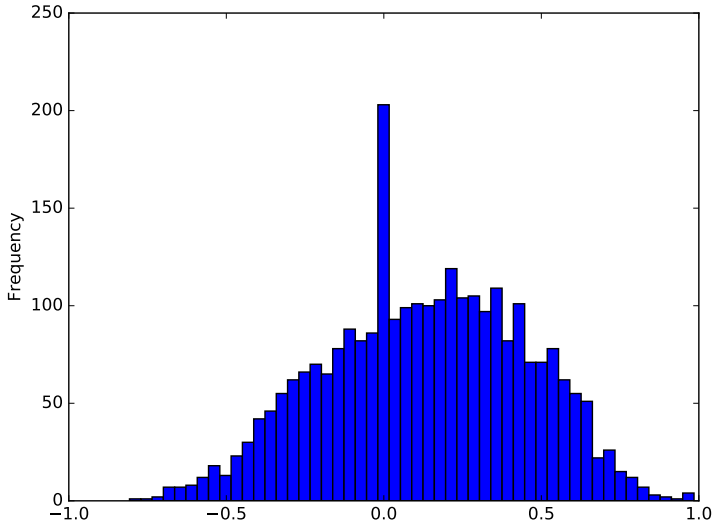

Figure 8: dt/D Histogram for Stations S506 and S501.

this help to show that the software is operating correctly?

I have not found anything in the HiSPARC documentation mentioning this problem, other than a short cryptic comment in Fokkema's PhD thesis describing the hardware and expected results from the HiSPARC experiment (Fokkema 2012) referring to an unexplained systematic time offset that he found when comparing a HiSPARC detector station with data from a professional detector.

4 Alternative Lines of Investigation

For coincidences registered in the Science Park, for which we can calculate 3-station direction solutions, it will also often be the case that there are accurate direction solutions using just the timing data between the four-plates of each of these detections. In principle we might compare the multi-station solutions with the supposedly more accurate individual station direction calculations.

This is feasible because the data for a multi-station coincidence contains a data reference to the individual station event stored in a separate table.

References

Fokkema, D. (2012), The HiSPARC Experiment, PhD thesis, University Twente, The Netherlands.