D.B.R.A. Fokkema

HiSPARC

# 1 Inleiding

Het HiSPARC project verzamelt al jaren data van tientallen stations in voornamelijk Nederland, Denemarken en Engeland. Het is gebruikelijk in de wetenschap dat het analyseren van data niet gebeurt in kant-en-klare programma's, zoals een spreadsheet-programma. Een programma als Excel wordt vaak gebruikt om de cijfers van een klas in te voeren of een balans bij te houden. In het bedrijfsleven wordt het ook vaak (mis/ge)bruikt om met veel ingewikkelde formules een analyse van bedrijfsgegevens uit te voeren. Hierbij gaat het bijna altijd om relatief weinig gegevens. Een dag HiSPARC data, daarentegen, bestaat met gemak uit 50 000 tot 60 000 regels.

In de wetenschap weet je nooit wat je kunt verwachten als je onderzoek doet. Een uitgebreide analyse van gegevens wordt daarom normaal gesproken geprogrammeerd. Een programmeeromgeving als *Mathematica*, of een programmeertaal als *Python*, behoort tot het standaardgereedschap van de onderzoeker. Het aanleren en zelf schrijven van zo'n analyse kost alleen heel veel tijd.

Onderzoek wordt vrijwel altijd gedaan in een onderzoeksgroep, met meerdere wetenschappers. Vaak schrijven één of twee wetenschappers de grote lijnen van een analyse en worden deze programma's gebruikt door de overige leden van de groep.

# 2 Datasets downloaden en bekijken

Het HiSPARC-team heeft een relatief gebruiksvriendelijk programma geschreven dat beschikbaar is in de webbrowser: de *data retrieval tool*[1]. Met dit programma kun je alle HiSPARC data downloaden en eenvoudige tot enigszins complexe analyses uitvoeren.

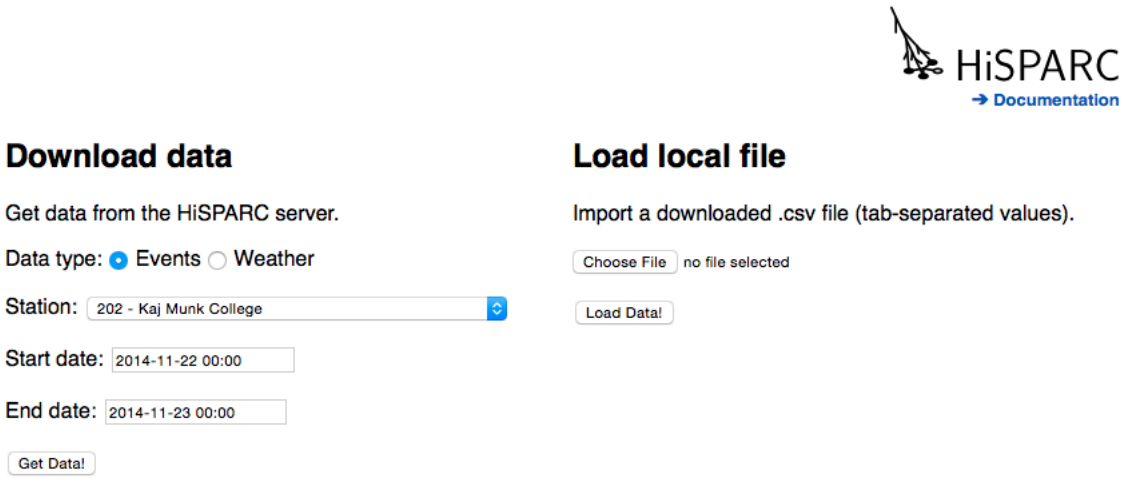
**Opdracht 1:** Open een browser en ga naar `http://data.hisparc.nl/media/jsparc/data_retrieval.html`. Controleer dat je pagina overeenkomt met Figuur 2.1. Bekijk de pagina goed.

In dit werkblad richten we ons op de linkerkolom: *download data*. Hier is het mogelijk om data van de HiSPARC servers te downloaden in de browser om vervolgens de data te analyseren.

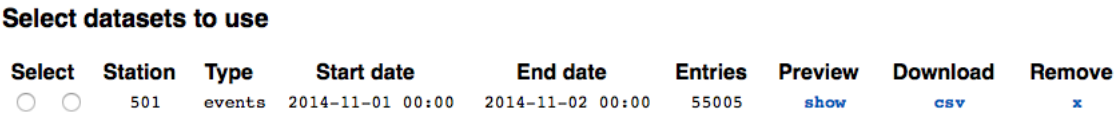
**Opdracht 2:** We gaan in eerste instantie *events* downloaden. Kies station 501 (Nikhef) en een startdatum van 1 november 2014, en een einddatum van 2 november 2014. Klik op *Get Data!*.

Tijdens het downloaden wordt het HiSPARC logo rechtsbovenaan de pagina geanimeerd. De animatie bootst een shower na die op het aardoppervlak wordt gedetecteerd. Als de animatie stopt is de dataset gedownload.

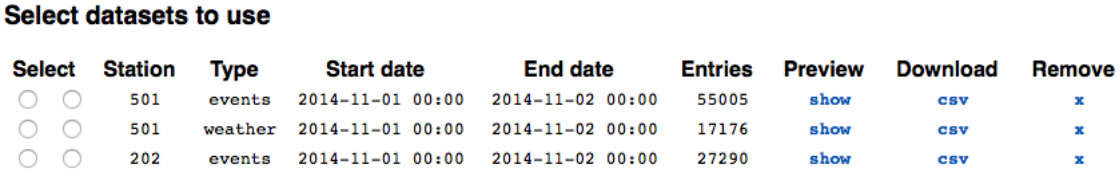
**Opdracht 3:** Bekijk Figuur 2.2. Dit kopje verschijnt op de pagina zodra een dataset beschikbaar is. Download voor hetzelfde station en dezelfde data ook eens meteorologische gegevens (*data type: weather*). Download ook eens events voor station 202. Het kopje datasets komt nu overeen met Figuur 2.3.



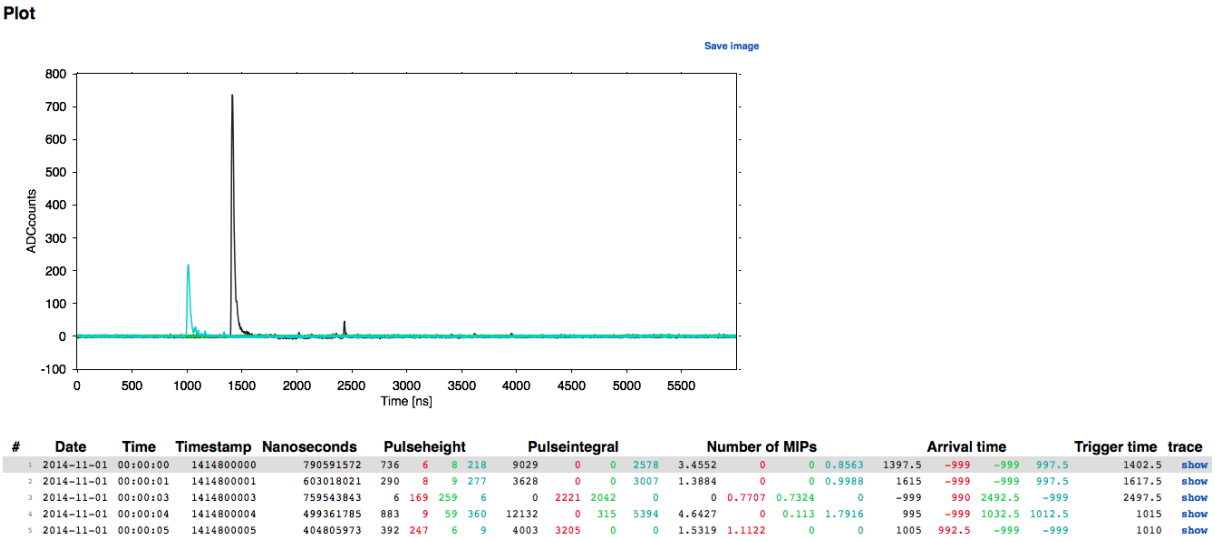

**Figuur 2.1** – Openingspagina van de data retrieval tool.




**Figuur 2.2** – Na downloaden is er een dataset beschikbaar.




**Figuur 2.3** – Meerdere datasets kunnen na elkaar worden gedownload.




**Figuur 2.4** – Preview van een HiSPARC event.

Bekijk Figuur 2.3. Het getal onder *entries* is het aantal metingen (events of meteorologische gegevens) dat beschikbaar is. Als je klikt op *tsv* onder *download*, dan download je alle data als tekstbestand. Je kunt de gegevens dan inladen in een ander programma. Om een dataset uit het geheugen te verwijderen klik je op het kruisje onder *remove*.

Een overzicht van de ruwe data is beschikbaar door te klikken op *show* onder *preview*.

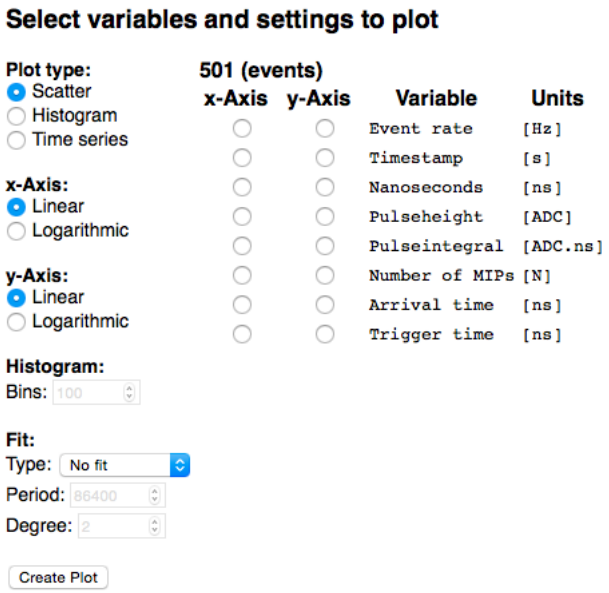
**Opdracht 4:** Bekijk een preview van de events van station 501. Klik voor het bovenste event op *show* onder het kopje *trace* helemaal rechts. Je scherm moet nu overeen komen met Figuur 2.4. De grafiek is een weergave van de ruwe data van een HiSPARC event. Geladen deeltjes zijn door de detectoren gegaan en hebben een lichtspoor nagelaten. Bekijk de grafiek en bestudeer de getallen van het eerste event in de tabel. Verklaar de getallen onder *pulseheights*, *arrival time* en *trigger time*.

De kolom *pulseintegral* geeft aan hoe groot de oppervlakte is onder het signaal. Dit is een maat voor hoeveel energie de deeltjes hebben achtergelaten in de detector en wordt gebruikt om een schatting te maken van het aantal geladen deeltjes dat door een detector ging. Deze schatting is weergegeven onder het kopje *number of MIPs*. MIP staat voor *minimum-ionizing particle*, ofwel een deeltje dat een minimale hoeveelheid energie verliest door ionisatie. Deeltjes in showers vallen in deze categorie.

In de tabel komt soms het getal -999 voor. Dit betekent dat het niet mogelijk was een voorlopige analyse uit te voeren op de data. In de praktijk betekent dit dat er in die detector geen deeltjes zijn gedetecteerd. Ook kan het getal -1 voorkomen. Dat betekent dat er voor die kolom geen data beschikbaar is, bijvoorbeeld als je data bekijkt van een station met twee detectoren, in plaats van vier.

**Opdracht 5:** Bekijk een preview van de data van station 202. Hoeveel detectoren heeft dit station?

[1] `http://data.hisparc.nl/media/jsparc/data_retrieval.html`




**Figuur 3.1** – Instellingen voor grafieken. Hier kun je kiezen welke gegevens je wilt weergeven en wat voor type grafiek je wilt krijgen.

# 3 Analyse van HiSPARC events

Om een dataset te analyseren moeten we die eerst selecteren.

**Opdracht 6:** Klik op het eerste bolletje van de dataset van events van station 501. Op de pagina verschijnt nu een stuk met instellingen om een grafiek te maken (Figuur 3.1).

## 3.1 Scatter plots

Standaard maak je een *scatter plot*. Dit soort grafieken maak je ook tijdens een practicum natuurkunde. Je vergelijkt twee gegevens (bijvoorbeeld kracht en uitrekking van een veer) en voor iedere meting zet je een punt in de grafiek. Het enige verschil met een 'gewoon' practicum is dat de HiSPARC dataset veel meer metingen bevat. In plaats van een paar punten met een lijn zul je nu vaak een wolk van punten zien.

Een scatter plot is ideaal om te onderzoeken of bepaalde grootheden van elkaar afhankelijk zijn (zoals kracht en uitrekking bij een veer). Als er een verband is, dan zie je dat in de grafiek. Is er *geen* verband, dan zie je slechts een wolk van punten.

**Opdracht 7:** Kies *timestamp* voor de x-as en óók *timestamp* voor de y-as. Klik dan op *create plot* om een grafiek te maken.

Als variabelen van elkaar afhankelijk zijn, zeggen we ook wel dat ze met elkaar *correleren*. De grafiek die je net gemaakt hebt is een extreem voorbeeld. De waardes op de x- en y-as zijn *exact* gelijk. Je kunt veel leren door correlaties te onderzoeken en te kijken naar de vorm die verschijnt in de grafieken. Als je nadenkt over *waarom* een correlatie een bepaalde vorm heeft, kan het helpen om de grafiek nog een keer te plotten, maar dan de variabelen voor de x- en y-as om te draaien.

**Opdracht 8:** Kies *timestamp* voor de x-as en *nanoseconds* voor de y-as. Klik dan op *create plot* om een grafiek te maken.

Dit is een extreem voorbeeld van *ongecorreleerde* variabelen. Er is geen samenhang. Dit komt omdat het tijdstip van een event op een dag (timestamp) niets te maken heeft met of het event aan het begin van een seconde, middenin een seconde, of aan het eind van een seconde wordt gemeten (nanoseconds).

**Opdracht 9:** Maak een scatter plot met pulshoogte op de x-as, en pulsintegraal op de y-as. Probeer de vorm te verklaren.

**Opdracht 10:** Draai het nu om: zet pulsintegraal op de x-as, en pulshoogte op de y-as. Probeer weer de vorm te verklaren. Zie het antwoord[2] in de voetnoot. Vind je dit duidelijker?

**Opdracht 11:** Onderzoek andere correlaties. Welke grootheid wordt *direct* gebruikt om een schatting te maken van het aantal MIPs in een detector?

## 3.2 Histogram

Een *histogram* is een grafiek waarin wordt aangegeven *hoe vaak* een bepaalde waarde voorkomt. Bijvoorbeeld: hoeveel leerlingen hebben een lengte tussen de 1,80 m en 1,85 m?

Als je als plottype *histogram* kiest, kun je geen y-as meer kiezen. De grafieken die gemaakt worden hebben een y-as die loopt van de kleinste tot de grootste waarde. *Hij begint vaak niet bij nul!*

**Opdracht 12:** Maak een histogram van de event rate (het aantal events per seconde). Welke event rate komt het meest voor? Hoe groot is de spreiding?

**Opdracht 13:** Maak een histogram van de timestamps (tijdstippen van de events). Als je één dag data hebt, zet het aantal bins dan op 24. In het histogram verschijnt dan het aantal events per uur.

**Opdracht 14:** Maak een histogram van de pulshoogten. Maak histogrammen voor 10, 50, 100, 500 en 1000 bins. Wat is het voordeel van veel bins? En wat is een nadeel?

[2] In deze grafiek zie je dat events met een grote integraal (pulsoppervlak), ook een grote hoogte hebben. In eerste instantie gaat dat gelijk op: als er meer deeltjes door een detector gaan dan wordt de puls hoger én breder. Maar als er heel veel deeltjes door een detector gaan, wordt de oppervlakte van de puls wel (flink) groter, maar niet heel veel meer hoger. De verklaring hiervoor is dat de elektronica pulsen met een grote spanning (de 'hoogte' van de puls) niet goed aankan.

Als er een groot verschil is tussen bins met veel counts en bins met weinig counts is het vaak moeilijk om te zien hoe de grafiek precies loopt. Je kunt dan kiezen voor een *logaritmische* schaalverdeling.

**Opdracht 15:** Gebruik een logaritmische y-as in een pulshoogtehistogram met 1000 bins. Geef een verklaring voor de bult in de grafiek.

### 3.3 Time series

Een *time series* grafiek is een scatter plot waarbij de tijd langs de x-as staat. Eigenlijk krijg je precies hetzelfde als je in een scatter plot kiest voor *timestamps* langs de x-as. Het enige verschil is dat de tijdstempels worden omgeschreven naar een leesbare weergave van datum/tijd.

### 3.4 Correlaties met weerdata

Bovenaan de pagina, onder het kopje *select datasets to use*, selecteer de weerdata als tweede dataset. Onder het kopje *select* staat dan in de eerste kolom een bolletje bij *events*, en in de tweede kolom een bolletje bij *weather*.

Je kunt nu vrij kiezen uit variabelen van events en weerdata om met elkaar te plotten. Om een verband (correlatie) tussen twee grootheden te onderzoeken gebruik je een scatter plot.

**Opdracht 16:** Onderzoek een mogelijk verband tussen de luchtdruk en de temperatuur.

**Opdracht 17:** Onderzoek een mogelijk verband tussen de luchtdruk en het aantal gedetecteerde showers per seconde.

Als je denkt een verband gevonden te hebben, kun je bij de grafiekopties onder het kopje *fit* de computer laten zoeken naar het precieze verband. Kies eens voor *linear* en maak een nieuwe grafiek. Onderaan de grafiek staat dan de functie van de beste lijn door de data. Beoordeel zelf in de grafiek of je vindt dat de computer een goed verband heeft gevonden. Vaak wordt een verband pas duidelijk als je veel data hebt verzameld. Eén dag is vaak niet genoeg. Kies dan liever een week of een maand.

**Opdracht 18:** Onderzoek nogmaals het verband tussen luchtdruk en event rate, en fit een lineaire functie. Zie je een duidelijk verband?

## 4 Kort eigen onderzoek

Voor dit eigen onderzoek kun je zelf data downloaden. Je mag zelf kiezen welke stations je gebruikt,[3] en hoeveel data je nodig hebt (een dag, een week, een maand, ...). Je kunt denken aan één van onderstaande opdrachten, maar je mag ook zelf iets verzinnen.

[3] Zie ook de kaart op de HiSPARC website: `http://data.hisparc.nl/show/stations_on_map/`. Als je ver inzoomed kun je de stationsnummers zien.

**Opdracht 19:** Onderzoek het verband tussen verschillende weergegevens: luchtdruk en temperatuur, temperatuur en hoeveelheid zonnestraling, windkracht en windrichting, enz.

**Opdracht 20:** Onderzoek het verband tussen het aantal waargenomen showers en weergegevens als luchtdruk, temperatuur, enz.

**Opdracht 21:** Onderzoek een mogelijke periodiciteit in aantallen waargenomen showers: zie je bijvoorbeeld overdag meer showers dan 's nachts?

**Opdracht 22:** Google naar *space weather* o.i.d. en onderzoek of je meer of minder showers ziet als er zonneuitbarstingen zijn.

**Opdracht 23:** Ga op zoek naar dagen met onweer (buienradar?). Bekijk de HiSPARC data voor verschillende stations, op het moment dat de onweersbuien over dreven. Je kunt denken aan aantal events per seconde, gemiddelde pulsintegraal, maar ook luchtdruk en temperatuur.