

COAR - Logboek

notitie 04 januari 2016, Henk van den Berg

Een deel van de archeologische datasets heeft geen coördinaten in de metadata. Op zoek naar ontbrekende geo-locatie van datasets werden in de zomer van 2015 alle pdf-bestanden uit archeologische datasets gescand. Een deel van deze datasets had al coördinaten in de descriptive metadata, waarmee we een testgroep hadden om de betrouwbaarheid van het extractieproces te monitoren.

Dit verslag is een logboek van de analyse van de verzamelde data – meer dan dat is het niet.

Eerst kijken we naar de betrouwbaarheid van de gevonden resultaten, door bij datasets, waarvan de geo-locatie bekend is uit de metadata, de coördinaten gevonden in pdf-bestanden te vergelijken.

Vervolgens proberen we voor de datasets met onbekende geo-locatie per dataset een centraal punt te extrapoleren en een gradatie van betrouwbaarheid te berekenen.

De meest onbetrouwbare resultaten uit de lijst met ontbrekende coördinaten moeten nog gecontroleerd.

Enkele datasets hebben een grote spreiding in geografische locatie (verzamelrapporten bijvoorbeeld). Soms is het niet zinvol deze datasets met één centraal punt te markeren.

Locatie software:

<https://github.com/dans-er/coar>
<https://github.com/dans-er/coar-ui>
https://github.com/dans-er/coar_an

Verslagen en data:

<https://surfdrive.surf.nl/files/index.php/apps/files/?dir=%2FbestandsherkenningEASY%2Fcoar>

1 Distributie

Over hoeveel datasets gaat het eigenlijk?

```
SELECT count(*) FROM coar_n.tdatasets;
```

24.935 datasets werden onderzocht

```
SELECT count(*) FROM coar_n.tdatasets  
WHERE co_emd > 0;
```

18.547 datasets met een of meer coordinaten in descriptive metadata (EMD)

```
SELECT count(*) FROM coar_n.tdatasets  
WHERE co_emd = 0;
```

6.388 datasets met geen coordinaten.

Het is een *long-tail*-distributie: veel datasets met 0 of 1 coordinaat, weinig datasets met veel coordinaten.

coordinaten in EMD	aantal datasets	aantal coordinaten
0	6.388	-
1	16.825	16.825
2	921	1.842
3	130	390
4	533	2.132
5	43	215
6	29	174
7	6	42
8	18	144
9	3	27
10	11	110
11	3	33
12	7	84
14	2	28
15	1	15
16	3	48
17	2	34
18	2	36
19	1	19
24	1	24
25	2	50
31	1	31
38	1	38
47	1	47
51	1	51
	24.935	22.439

tabel 1. distributie van coordinaten uit metadata over datasets.

2 Vinden van coordinaten

Methode-beschrijving

Bij vinden van coordinaten in pdf's maakten we gebruik van *regular expressions*. De parser in Apache Tika gooit per blok tekst een *line*. Blokken tekst kunnen van grootte verschillen, soms heeft een blok de grootte van een regel in de zichtbare

tekst, soms is het een woord of zelfs een deel van een woord, soms een hele paragraaf. De gebruikte patterns vinden coordinaten in die blokken aan de hand van de trefwoorden, zoals `co??ord*`, windrichtingen of de letters `x` en `y`. Soms was het nodig een alert te zetten om in de volgende blok tekst naar coordinaten te zoeken.

Zie: <https://github.com/dans-er/coar/blob/master/src/main/java/nl/knaw/dans/coar/walk/CoordinateDetector.java>

Het vinden van patterns werd gefailliet met een testset van 53 willekeurige pdf's. https://surfdrive.surf.nl/files/index.php/apps/files?dir=/bestandsherkenningEASY/coar/learning_set (die links naar SurfDrive doen 't maar half!)

De patterns, combinaties van patterns + overige logica leverden vijf methoden van vinden van coordinaten. De methode van vinden, aangegeven met een nummer, werd, samen met de coordinaten opgeslagen in de database.

methode	aantal	omschrijving
0	22.423	coordinaten in EMD
1	17.924	<code>thePattern (co??ord*: *)</code>
2	236.713	cummulatiefA (in volgende blok)
3	7.957	cummulatiefB (in volgende blok)
6	48.271	<code>xyPattern (X: *, Y:*)</code>
110	1.432	<code>nozwPattern (windrichtingen Noord, Zuid, Oost, West)</code>
334.720		

tabel 2. Aantallen coordinaten die met verschillende methoden werden gevonden.

De coordinaten uit metadata werden in dezelfde tabel opgeslagen als die gevonden in pdf's. Methode 0 is geen methode zoals hierboven beschreven, maar geeft aan dat de coordinaten uit EMD afkomstig zijn. Boxes in EMD werden vertaald naar punt opgeslagen. Het verschil in aantal met *totaal coordinaten* uit tabel 1 ($22.439 - 22.423 = 16$) wordt verklaard doordat punt- en vertaalde box-coordinaat bij sommige datasets exact samenvallen. We verzamelden unieke punten.

3 Valideren tegen controlegroep

Door de gevonden coordinaten per methode te vergelijken met de bekende coordinaten uit EMD krijgen we een idee van de betrouwbaarheid van de verschillende methodes. Die vergelijking kan natuurlijk op verschillende manieren. Hieronder wordt de methode gevuld, waarbij het rekenkundig gemiddelde en de standaard deviatie voor de uit EMD bekende punten voor de x- en y-component apart berekend worden en vergeleken met de x- en y-waarden van gevonden punten, rekening houdend met die standaard deviatie voor x- en y-coordinaat van het controlepunt. Voordeel van sd apart op x en y is dat je bij datasets met meerdere coordinaten in de EMD de vorm van het gebied betreft. Er zijn niet veel datasets met meer dan 1 coordinaat (1.708 van de 18.547). Bovendien zijn de boxes uit EMD platgeslagen tot singuliere punten.

Stappen:

1. Van de 18.547 datasets met bekende coordinaten, bereken het rekenkundig gemiddelde van x- en y-coordinaten en de standaard deviatie (sd) voor x en y. Als er maar een stel coordinaten is, stel de waarde van de sd's op 0. De 18.547 datasets met bekende coordinaten zijn samen de controlegroep.
2. Bereken voor alle gevonden punten uit de controlegroep de afstand van x en y tot het berekende rekenkundig gemiddelde en betrek daar de berekende sd voor x en y bij. De controlegroep heeft 231.655 gevonden punten; in totaal werden er 312.297 punten gevonden in pdf's.

Voor details zie script: https://github.com/dans-er/coar_an/blob/master/validate.R

3.1 Puntenwolk

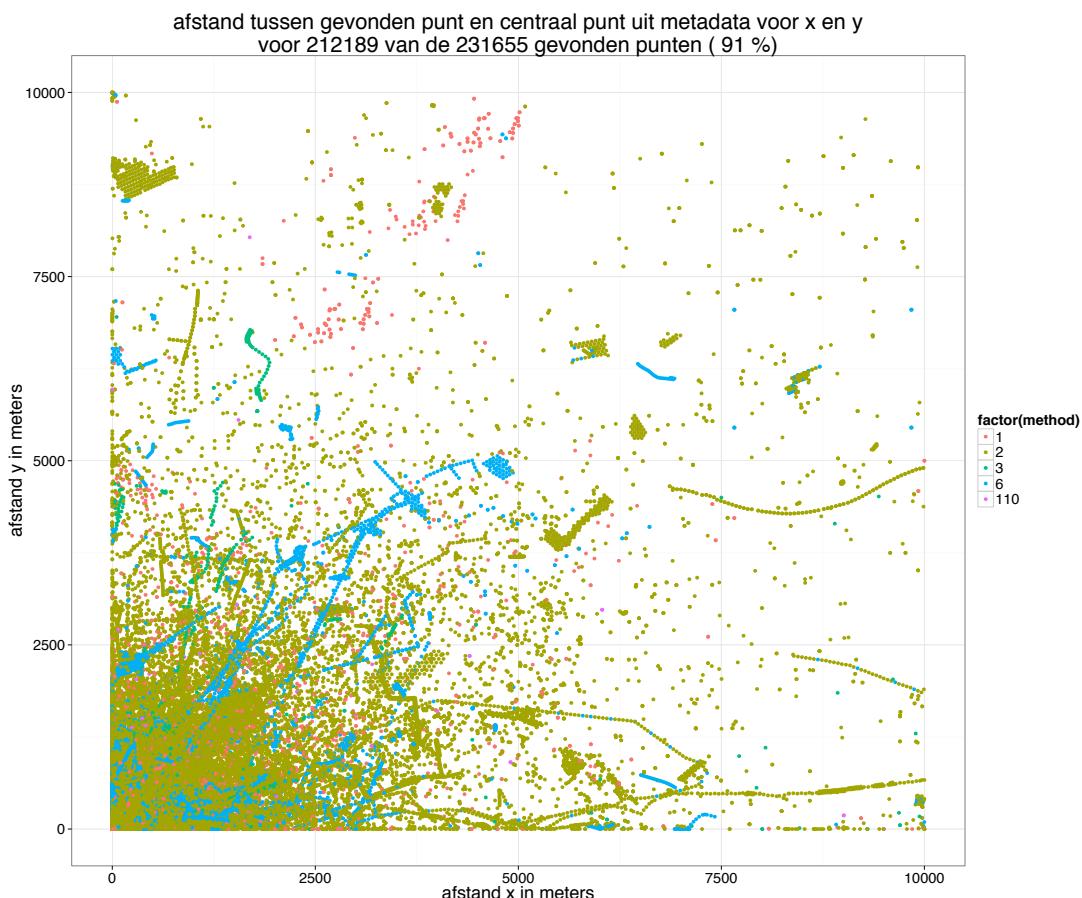


fig. 1. Afstand van de gevonden punten tot het centrale punt uit de metadata. Gegevens gefilterd tot maximaal 10 km. afstand. Patronen, lijnen en slierten zijn waarschijnlijk boorputten.
https://github.com/dans-er/coar_an/blob/master/images/puntenwolk.pdf

In figuur 1 is de afstand over x en y van gevonden punten geplot ten opzichte van het berekende centrale punt, rekening houdend met de standaard deviatie voor x en y. Extreme afstanden (meer dan 10 km.) zijn weggelaten, om te zorgen dat de schaal van de grafiek verhindert dat je iets ziet. Regelmaat en patronen duiden waarschijnlijk op reeksen boorputten uit een onderzoek.

De meeste punten zijn binnen 2,5 km. van het centrale punt. Het volgende scenario kan natuurlijk spelen. Er is een onderzoeksgebied van 5 x 5 km., maar in de metadata van de dataset is maar 1 centraal punt geregistreerd. In de pdf worden

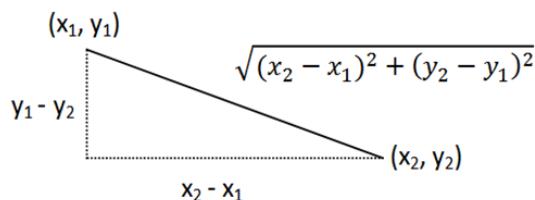
coordinaten van boorputten gevonden, weliswaar *binnen* het onderzoeksgebied, maar nog steeds met een afstand van 2,5 km tot het centrale punt. Er is in dit geval geen standaard deviatie voor x en y die de afstand van 2,5 km. kan compenseren, omdat er maar 1 punt is, het centrale punt. En hetzelfde kun je natuurlijk zeggen voor datasets met een onderzoeksgebied van 10 of 100 km. en 1 punt in de metadata. Dit verstoort dus de meting die we hier aan 't doen zijn: datasets met bekende punten gebruiken om de mate van correctheid van gevonden punten te controleren.

Ook de correctheid van de coordinaten in EMD is natuurlijk van belang. Als die fout zijn, dan zegt de vergelijking tussen bekend punt (maar fout punt) en gevonden punt helemaal niks. Bij de dataset met de grootste afwijking tussen gevonden punten en centraal punt is dat meteen raak: de coordinaten uit EMD liggen ergens in Noord Frankrijk en onder Bremen, de gevonden punten zijn wel correct, in de buurt van Doorn. <http://zandbak11.dans.knaw.nl/shiny/coar/dstoon-app/?id=easy-dataset:26122>

Overigens staat de dataset uit de alinea hierboven in state 'draft'. Die niet gepubliceerde datasets zou je uit de controlegroep willen filteren, maar de state van een dataset uit de administratieve metadata hebben we niet in de database.

3.2 Afstand tot controlepunt

We brengen de afzonderlijke afstanden tussen x en y terug tot 1 waarde, gewoon de afstand tussen het gevonden punt en het centrale punt uit de metadata. De afstand tussen twee punten in een plat coördinatenstelsel is gelijk aan de vierkantswortel van de som van het kwadraat van het verschil tussen x- en y-waarden van die twee punten. We noemen de gevonden waarde *distance*.



> summary(dt\$distance)						
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	
0	39	151	18270	755	455700000	

De grootste afstand tussen gevonden punt en controlepunt is 455.700 km. Het controle punt staat als

RD (in m.) X: 54530.343 Y: 452868.486

in de metadata, dus met een punt als decimaal scheidingsteken (en kennelijk milimeter-precisie). Of die notatie correct is, laat ik aan Valentijn over; de berekening heeft er moeite mee en plaatst de punt onder India, na zo'n 10 keer de aarde te zijn rondgetrokken. De gevonden punten liggen correct op de Noordzee. <http://zandbak11.dans.knaw.nl/shiny/coar/dstoon-app/?id=easy-dataset:58904>

De summary laat ook nog zien dat de helft van de gevonden punten op 151 meter of minder van het controlepunt ligt, $\frac{3}{4}$ van de gevonden punten ligt op 755 meter of minder van het controlepunt en $\frac{1}{4}$ op 39 meter of minder van het controlepunt.

Voor verschillende methoden van vinden kunnen we ook quantielen berekenen en daarmee de mate van betrouwbaarheid van de methode.

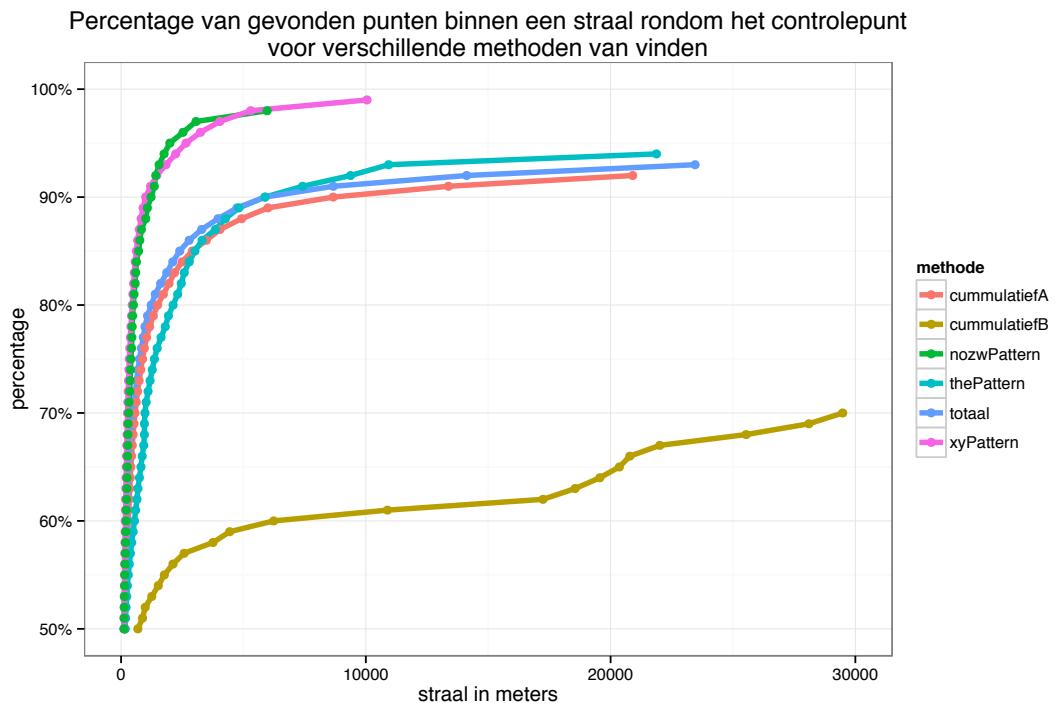


fig. 2. Welk percentage van gevonden punten nog binnen een bepaalde straal rondom het controlepunt ligt, geeft de mate van betrouwbaarheid van de gebruikte methode. Hoe steiler de lijn, hoe betrouwbaarder de methode.

https://github.com/dans-er/coar_an/blob/master/images/distance.pdf

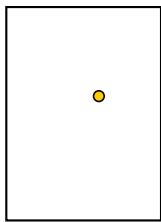
In figuur 2 zijn de percentages van gevonden punten getekend, tegen de afstand in meters van het controlepunt, tot een straal van 30 km. rondom het controlepunt. Met enige slagen om de arm: de methodes nowzPattern en xyPattern komen als beste uit de bus; de methode met de naam cummulatiefB scoort het slechtst. Met uitzondering van cummulatiefB geldt voor alle methoden: 90% van de gevonden punten ligt binnen een straal van 10 km. rondom het controlepunt.

3.3 Samenvatting validatie tegen controlegroep

De controlegroep is niet helemaal zuiver. Er zitten ook datasets in die in status *draft* of *submitted* staan en nog niet zijn gecontroleerd door archivarissen. Hoe groot dit aandeel is, weten we niet want de status van datasets is niet naar database gehaald.

Aan de lijnen en patronen in figuur 1 kunnen we zien dat ook punten op afstanden van 10 km. of meer tot het controlepunt nog best valide coordinaten uit het onderzoeksgebied kunnen zijn. Als tenminste onze veronderstelling juist is, dat de lijnen en patronen de coordinaten van aaneengesloten series boorputten laten zien.

Met uitzondering van één methode kunnen we zeggen dat 90% van de gevonden punten binnen een straal van 10 km. rond het controlepunt ligt. En dat de punten uit die groep *naar alle waarschijnlijkheid** correct opgespoorde coordinaten uit de content van de pdf-bestanden zijn.



- *Naar alle waarschijnlijkheid:* Gegevens werden gefilterd op $10.000 \leq x \leq 300.000$ en $300.000 \leq y \leq 700.000 \Rightarrow 290 * 400 \text{ km}^2$. geeft een gebied van 116.000 km^2 . De kans dat een punt toevallig in een gebied binnen een straal van 10 km. rondom controlepunt komt is kleiner dan 0,3%.

4 Datasets met geen coordinaten in EMD

```
SELECT count(*) FROM coar_n.tdatasets  
WHERE co_emd = 0;
```

6.388 datasets met geen coordinaten in EMD,

```
SELECT count(DISTINCT datasetId) FROM tspatials  
JOIN tdatasets ON datasetId = parent_datasetId  
WHERE co_emd = 0;
```

4.044 datasets met geen coordinaten in EMD, wel coordinaten in pdf. (63%)

```
SELECT count(*) FROM tspatials  
JOIN tdatasets ON datasetId = parent_datasetId  
WHERE co_emd = 0;
```

80.642 coordinaten gevonden, gemiddeld 20 coordinaten per dataset.

2 manieren om de betrouwbaarheid van gevonden punten te controleren. Code in https://github.com/dans-er/coar_an/blob/master/spread.R

4.1 Blind varen: spreiding

Als er meerdere coordinaten in pdf's zijn gevonden en die punten liggen betrekkelijk dicht bij elkaar, dan kunnen we, onder voorbehoud, stellen dat de gevonden punten correct zijn. Tegenovergesteld hoeft natuurlijk niet. Punten verspreid over heel Nederland kunnen nog correct gelezen coordinaten zijn. Vgl. <http://zandbak11.dans.knaw.nl/shiny/coar/dstoon-app/?id=easy-dataset:15403> een verzamelrapport.

Als de spreiding van de puntenwolk erg groot is, dan kan dat komen door foutief gelezen punten gevonden met minder betrouwbare methoden.

Verkleinen puntenwolk. for each dataset:

1. bereken de standaarddeviatie voor x-component en y-component van gevonden punten als `sd_x` en `sd_y`.
2. bereken spreiding: `distance:=as.integer(sqrt(sd_x^2 + sd_y^2))`
3. Is de spreiding groter dan 10 km. en er zijn nog punten gevonden met andere methoden over:
 - a. laat punten verkregen met minst betrouwbare methode uit de berekening en begin weer bij 1.

Op deze manier zijn dubieuze punten gevonden met methode 3 (cumulatiefA) en 2 (cumulatiefB) gemerkt.

4.2 Blind varen: geonames

De betrouwbaarheid van de puntenwolk wordt door twee dingen bepaald: het aantal punten in de puntenwolk en de spreiding. Als we 48 punten vinden en al die punten liggen niet verder dan 2 km. uit elkaar, dan weten we vrij zeker dat die punten correct zijn. Hebben we weinig punten, bijvoorbeeld 2 per dataset, of liggen de punten ver uit elkaar, dan is de betrouwbaarheid niet groot. Hoe controleren we die kritische punten? Antwoord: geonames.

Veel datasets hebben plaats- of gemeentenamen in hun titel. De API-call findNearbyPostalCodes van geonames geeft voor het opgegeven punt o.a. gemeentenaam en plaatsnaam. Vgl.:

<http://api.geonames.org/findNearbyPostalCodes?lat=52.651907&lng=5.772913&username=genda>

Als de op deze manier gevonden plaats- en/of gemeentenamen terugkomen in de titel dan is dit een sterke aanwijzing dat de gevonden punt correct is.

Geonames stelt grenzen aan het aantal requests (max. 2000 requests per uur). Namen opvragen voor alle 80.000 gevonden punten zou te lang duren. Punten werden opgevraagd voor kritische records. 15.982 van de 80.642 records werden op deze manier van geo-namen voorzien. Definitie: datasets met punten waarvoor geo-namen zijn gevonden hebben geo-evidence.

4.3 Gewicht toekennen

Met het voorwerk uit 4.1 en 4.2 kunnen we de gevonden coordinaten van een gewicht voorzien. Alle records krijgen een basiswaardering (weight) van 1.

- Datasets zonder geo-evidence. In dit geval vertrouwen we alleen op de puntenwolk.
 - Een onder 4.1 gevonden dubieuze punt die sterk bijdraagt aan het vergroten van de spreiding krijgt waardering 0, doet niet meer mee.
- Datasets met geo-evidence.
 - Voor het record zelf is geen geo-evidence en onder 4.1 gemerkt als dubieuze punt: waardering 0, doet niet meer mee.
 - Voor het record is een *name* gevonden: verhoog waardering met 5 punten.
 - voor het record is een *adminName2* (gemeentenaam) gevonden: verhoog waardering met 4 punten.

Als we records met 0 gewicht schrappen, dan houden we nog 77.471 van de 80.642 records over.

4.4 Aggregatie

In deze stap berekenen we a.) een benadering van de geografische locatie van de dataset, één coördinaat per dataset en b.) een aantal variabelen die mogelijkwijs iets zeggen over de nauwkeurigheid en de betrouwbaarheid van de berekende

coördinaat. In `SURFdrive/bestandsherkenningEASY/coar/results/data/aggregate.csv` het resultaat van deze berekeningen. Deze gegevens zijn weer geïmporteerd in Excel en verder bewerkt: `SURFdrive/bestandsherkenningEASY/coar/results/data/aggregate.xlsx`

Per kolom in de spreadsheet geven we hier een beschrijving.

a: datasetId

datasetId

b: dataset_id

link naar shiny pagina over de dataset. De berekening van de median coördinaat, 'center.pdf', is anders dan de hier gehanteerde.

c: emd_title

eerste dc:title uit EMD.

d: n_pdf

Het aantal coördinaten gevonden in pdf-bestanden.

e: n_used

Het aantal coördinaten gebruikt voor berekening van de centrale coördinaat na filteren als onder 4.3 gewicht toekennen

f: sdx

standaard deviatie van de x-coördinaat of 0 las er maar één waarde is:
`sdx = as.integer(ifelse(is.na(sd(coor_x)), 0, (sd(coor_x))))`

g: sdy

standaard deviatie van de y-coördinaat of 0 las er maar één waarde is:
`sdy = as.integer(ifelse(is.na(sd(coor_y)), 0, (sd(coor_y))))`

h: sdxw

standaard deviatie van gewogen x-coördinaat:
`sdxw = sd(rep(coor_x, times=weight))`

i: sdyw

zelfde als sdxw voor y-coördinaat

j: dist1

de diagonaal-lengte van de standaard deviatie voor x en y berekend over alle gevonden coördinaten: `dist1:=as.integer(sqrt((sd(coor_x))^2 + (sd(coor_y))^2))`

k: dist

de diagonaal-lengte van de standaard deviatie voor x en y berekend over gevonden punten na filtering: `dist:=as.integer(sqrt((sdx)^2 + (sdy)^2))`

l: distw

de diagonaal-lengte van de gewogen standaard deviatie:
`distw:=as.integer(sqrt((sdwx)^2 + (sdwy)^2))`

m: spread

maat voor de spreiding van punten voor datasets zonder geo-evidence
gewogen diagonaal-lengte gedeeld door aantal gebruikte punten: `spread=distw/n_used`

n: value

de som van het aan de punten toegekende gewicht: `value = sum(weight)`

o-r: minx, maxx, miny, maxy

de minimale en maximale waarde voor x en y van de gebruikte coördinaten.

s: opp

de oppervlakte van de gevonden coordinaten in km² beschreven als rechthoek:
opp = ((max(coor_x) - min(coor_x))/1000) * ((max(coor_y) - min(coor_y))/1000)

t-u: medianx, mediany

de gewogen median van x en y coordinaten:
medianx = as.integer(median(rep(coor_x, times=weight)))

v-w: plaats, gemeente

de met geonames gevonden name en adminName2 die terugkomen in de titel van de dataset.

x-y: lat, lon

naar WGS84 geconverteerde punt.

z-ab: google, os_map, geo_name

links naar google maps, openstreetmap en geonames voor de geconverteerde punt.

4.5 Handmatige controle

De dubieuze gevallen onder de berekende punten moeten nog gecontroleerd worden. Een begin is al gemaakt, zie kolommen corr_x, corr_y, controle en opmerking in het spreadsheet. Wat zijn dubieuze gevallen? Door het spreadsheet te sorteren op de waarde in verschillende kolommen, zouden dubieuze gevallen bovenaan moeten komen staan. Ik zou in ieder geval controleren:

1. value, ascending. Als er maar één of enkele punten in de pdf's zijn gevonden dan geldt de wet van de puntenwolk in een klein gebied niet. Vooral datasets waarvan de dist1 afwijkt van distw zijn verdacht (dan zijn er wel verschillende punten gevonden, waarvan sommige zijn gediskwalificeerd). Als er geo-evidence voor punten is, dan neemt value toe. Controleren tot value is 3 of 4, nog hoger?
2. spread, descending. Grote spreiding van puntenwolk kan mogelijk duiden op foute lezingen. Maar kan ook betekenen dat er verschillende opgravingen worden behandeld, of dat de dataset bijvoorbeeld twee rapporten bevat. In het laatste geval heeft de median, het centrale punt tussen de twee rapporten weinig zin want daar is niks.

Laatste opmerkingen:

- de set van datasets bevat ook datastes in state draft.
- er zijn pdf-rapporten waarin coordinaten volgens een ander stelsel. Controleer rapporten met niet-nederlandse titels?