Subject: Re: Duplicates - a new twist
Posted by Chris Lee on Tue, 18 May 2004 07:59:30 GMT
View Forum Message <> Reply to Message

In article <d33d6a4b.0405171324.1272c4e0@posting.google.com>, "Martin Doyle" <m.doyle@uea.ac.uk> wrote:

> Hello all,
> I have a problem which I've searched everywhere to try and solve...many
> posters on this newsgroup have had _similar_ problems but the
> resolutions didn't help me...anyway, here goes; I have a dataset which
> consists of 3 columns: longitude, latitude and a value for an emission
> ....
> What I need to do is to look through the dataset and sum the emissions
> when the coordinate is the same, resulting in a dataset with unique
> coordinates and a total emission for each grid point. Does anyone have
> ...
> Martin

Hi,

;data is fltarr(3,n)

;find possible collisions
d=sqrt(data[0,*]^2+data[1,*]^2)
threshold=1.0 ; numbers after the decimal point.
d=round(d*10^threshold)/10^threshold

;apply the single vector functions here, the result is a *possible* list
of collisions, since any point on a circle centred on the origin will
match, hopefully since your dealing with European data you won't collide
with anything in the Southern atlantic :)

There are probably better ways.

Chris.

Subject: Re: Duplicates - a new twist
Posted by Chris[1] on Tue, 18 May 2004 11:42:32 GMT
View Forum Message <> Reply to Message

The easiest way is to take copies of the lats & lons, reduce them to the resolution you think is sufficient that the same station has the same coordinate (independent of which country reports it); sort on one (say latitude), and then look for unique values, using uniq() on the sorted values. Then use the output from uniq() to look at whether points with the

same latitude also have the same longitude. It's an exercise in indexing :)

If you want a more robust technique - one that doesn't fall apart near the poles or the dateline, for example - use a spherical to Cartesian coordinate conversion, and do similarly, except now with the three coordinates.

Cheers;
Chris

"Martin Doyle" <m.doyle@uea.ac.uk> wrote in message
news:d33d6a4b.0405171324.1272c4e0@posting.google.com...
> Hello all,
>
> I have a problem which I've searched everywhere to try and
> solve...many posters on this newsgroup have had _similar_ problems but
> the resolutions didn't help me...anyway, here goes;
>
> I have a dataset which consists of 3 columns: longitude, latitude and
> a value for an emission of an air pollutant. European countries report
> the emission of this pollutant for the latitude longitude coordinates
> which are within their countries. However, some of the latitude,
> longitude coordinates lie on the borders of countries and therefore an
> emission is sometimes reported by 2 or more countries for the same
> coordinate (i,e. There are multiple instances of the same coordinate
> within the dataset).
>
> What I need to do is to look through the dataset and sum the emissions
> when the coordinate is the same, resulting in a dataset with unique
> coordinates and a total emission for each grid point.
>
> Does anyone have any ideas about how to go about this? I've seen posts
> on this newsgroup which have had problems with duplicate values in one
> column of data, but I'm unsure about how to go about it when there are
> 2 columns which need to be examined.
>
> Thanks guys...
>
> All the best,
>
> Martin

Subject: Re: Duplicates - a new twist
Posted by btt on Tue, 18 May 2004 12:32:45 GMT
View Forum Message <> Reply to Message

Martin Doyle wrote:

> I have a dataset which consists of 3 columns: longitude, latitude and
> a value for an emission of an air pollutant. European countries report
> the emission of this pollutant for the latitude longitude coordinates
> which are within their countries. However, some of the latitude,
> longitude coordinates lie on the borders of countries and therefore an
> emission is sometimes reported by 2 or more countries for the same
> coordinate (i,e. There are multiple instances of the same coordinate
> within the dataset).
>
> What I need to do is to look through the dataset and sum the emissions
> when the coordinate is the same, resulting in a dataset with unique
> coordinates and a total emission for each grid point.
>
> Does anyone have any ideas about how to go about this? I've seen posts
> on this newsgroup which have had problems with duplicate values in one
> column of data, but I'm unsure about how to go about it when there are
> 2 columns which need to be examined.
>

Hello,

You should consider using GRID_INPUT. This is from the docs...

The GRID_INPUT procedure preprocesses and sorts two-dimensional scattered data points, and removes duplicate values.

Ben

Subject: Re: Duplicates - a new twist
Posted by Bruce Bowler on Tue, 18 May 2004 19:10:28 GMT
View Forum Message <> Reply to Message

On Tue, 18 May 2004 08:32:45 -0400, Ben Tupper put fingers to keyboard and said:

> Martin Doyle wrote:
>
>> I have a dataset which consists of 3 columns: longitude, latitude and
>> a value for an emission of an air pollutant. European countries report
>> the emission of this pollutant for the latitude longitude coordinates
>> which are within their countries. However, some of the latitude,
>> longitude coordinates lie on the borders of countries and therefore an
>> emission is sometimes reported by 2 or more countries for the same
>> coordinate (i,e. There are multiple instances of the same coordinate

>> within the dataset).
>>
>> What I need to do is to look through the dataset and sum the emissions
>> when the coordinate is the same, resulting in a dataset with unique
>> coordinates and a total emission for each grid point.
>>
>> Does anyone have any ideas about how to go about this? I've seen posts
>> on this newsgroup which have had problems with duplicate values in one
>> column of data, but I'm unsure about how to go about it when there are
>> 2 columns which need to be examined.
>>
>
> Hello,
>
> You should consider using GRID_INPUT. This is from the docs...
>
>
> The GRID_INPUT procedure preprocesses and sorts two-dimensional
> scattered data points, and removes duplicate values.
>
> Ben

But Ben, he doesn't want to remove dup's, he wants to sum them...
(personally, I would have thought that average was better based on the
description, but what the heck...)

Bruce

--
+------------------+------------------------------------ ------------+
Bruce Bowler | What garlic is to salad, insanity is to art. -
1.207.633.9600 | Augustus Saint-Gaudens
bbowler@bigelow.org |
+------------------+------------------------------------ ------------+

Subject: Re: Duplicates - a new twist
Posted by btt on Tue, 18 May 2004 19:34:43 GMT
View Forum Message <> Reply to Message

Bruce Bowler wrote:

> On Tue, 18 May 2004 08:32:45 -0400, Ben Tupper put fingers to keyboard and
> said:
>
>
>> Martin Doyle wrote:
>>

>>
>>> I have a dataset which consists of 3 columns: longitude, latitude and
>>> a value for an emission of an air pollutant. European countries report
>>> the emission of this pollutant for the latitude longitude coordinates
>>> which are within their countries. However, some of the latitude,
>>> longitude coordinates lie on the borders of countries and therefore an
>>> emission is sometimes reported by 2 or more countries for the same
>>> coordinate (i,e. There are multiple instances of the same coordinate
>>> within the dataset).
>>>
>>> What I need to do is to look through the dataset and sum the emissions
>>> when the coordinate is the same, resulting in a dataset with unique
>>> coordinates and a total emission for each grid point.
>>>
>>> Does anyone have any ideas about how to go about this? I've seen posts
>>> on this newsgroup which have had problems with duplicate values in one
>>> column of data, but I'm unsure about how to go about it when there are
>>> 2 columns which need to be examined.
>>>
>>
>> Hello,
>>
>> You should consider using GRID_INPUT. This is from the docs...
>>
>>
>>   The GRID_INPUT procedure preprocesses and sorts two-dimensional
>> scattered data points, and removes duplicate values.
>>
>> Ben
>
>
> But Ben, he doesn't want to remove dup's, he wants to sum them...
> (personally, I would have thought that average was better based on the
> description, but what the heck...)
>

Awww! I was duped!

The DUPLICATES keyword for GRID_INPUT does everything BUT 'SUM'. Then again, setting DUPLICATES = 'all' should sort the data pairs so the duplicates are adjacent in the list. Then finding the pairwise difference between consecutive points should reveal where the duplicates are located. I have a vague memory of making a feature request for an INDEX output keyword that has the indices of the points retained by GRID_INPUT (relative to input vectors.) I remember getting a response at the time, but can't recall what it was... and obviously there is no such keyword in the current release.

Subject: Re: Duplicates - a new twist
Posted by R.G. Stockwell on Tue, 18 May 2004 19:35:25 GMT
View Forum Message <> Reply to Message

"Martin Doyle" <m.doyle@uea.ac.uk> wrote in message
news:d33d6a4b.0405171324.1272c4e0@posting.google.com...
> Hello all,
...
> which are within their countries. However, some of the latitude,
> longitude coordinates lie on the borders of countries and therefore an
> emission is sometimes reported by 2 or more countries for the same
> coordinate (i,e. There are multiple instances of the same coordinate
> within the dataset).

> What I need to do is to look through the dataset and sum the emissions
> when the coordinate is the same, resulting in a dataset with unique
> coordinates and a total emission for each grid point.

You could quickly make a one dimensional "index" array from the coordinates,
like coord = 1000*lat+lon , and use your one column uniq() and where()s.
Of course, handle the decimal points appropriately.
(or make it a string array of coordinates perhaps)

Offhand, it looks like you will need to loop through the uniq(coords) and
take the mean of the sum of where()d points.

Cheers,
bob