QGIS Application - Bug report #19517
Serious problem with rasters statistics calculated by QGIS with estimated option

2018-07-31 10:34 AM - Pedro Venâncio

Status:	Closed		
Priority:	High		
Assignee:	Matthias Kuhn		
Category:	Rasters		
Affected QGIS version:	3.3(master)	**Regression?:**	No
Operating System:	Windows 10	**Easy fix?:**	No
Pull Request or Patch supplied:	No	**Resolution:**	
Crashes QGIS or corrupts data:	Yes	**Copied to github as #:**	27345

Description

I think the problem described in these tickets can be more serious than just a visualization question: #11974, #14853 and #14835

Please try the following workflow.

1) Download this sample raster:
https://cld.pt/dl/download/de81b6ce-38a8-4539-be16-24a3e3ef72d1/perigosidade_int.tif

2) Run gdalinfo on it and see the output:

```
gdalinfo perigosidade_int.tif

Driver: GTiff/GeoTIFF
Files: perigosidade_int.tif
Size is 1039, 1701
Coordinate System is:
PROJCS["ETRS89 / Portugal TM06",
(...)
Origin = (73626.458811498901923,145193.972505581419682)
Pixel Size = (25.002694985400002,-25.002694985400002)
Metadata:
  AREA_OR_POINT=Area
Image Structure Metadata:
  INTERLEAVE=BAND
Corner Coordinates:
Upper Left  (   73626.459,  145193.973) (  7d15'30.17"W, 40d58'21.05"N)
Lower Left  (   73626.459,  102664.388) (  7d15'48.20"W, 40d35'22.48"N)
Upper Right (   99604.259,  145193.973) (  6d56'59.28"W, 40d58'11.13"N)
Lower Right (   99604.259,  102664.388) (  6d57'23.68"W, 40d35'12.70"N)
Center      (   86615.359,  123929.180) (  7d 6'25.36"W, 40d46'47.22"N)
Band 1 Block=1039x3 Type=Int16, ColorInterp=Gray
  NoData Value=32767
```

3) Then run gdalinfo with stats option, and the output is:

```
gdalinfo -stats perigosidade_int.tif

Driver: GTiff/GeoTIFF
Files: perigosidade_int.tif
```

```
Size is 1039, 1701
Coordinate System is:
PROJCS["ETRS89 / Portugal TM06",
(...)
Origin = (73626.458811498901923,145193.972505581419682)
Pixel Size = (25.002694985400002,-25.002694985400002)
Metadata:
  AREA_OR_POINT=Area
Image Structure Metadata:
  INTERLEAVE=BAND
Corner Coordinates:
Upper Left  (   73626.459,  145193.973) (  7d15'30.17"W, 40d58'21.05"N)
Lower Left  (   73626.459,  102664.388) (  7d15'48.20"W, 40d35'22.48"N)
Upper Right (   99604.259,  145193.973) (  6d56'59.28"W, 40d58'11.13"N)
Lower Right (   99604.259,  102664.388) (  6d57'23.68"W, 40d35'12.70"N)
Center      (   86615.359,  123929.180) (  7d 6'25.36"W, 40d46'47.22"N)
Band 1 Block=1039x3 Type=Int16, ColorInterp=Gray
  Minimum=2.000, Maximum=835.000, Mean=37.025, StdDev=71.526
  NoData Value=32767
  Metadata:
    STATISTICS_MAXIMUM=835
    STATISTICS_MEAN=37.024768992184
    STATISTICS_MINIMUM=2
    STATISTICS_STDDEV=71.52569727799
```

4) gdalinfo also creates a XML file with statistics info in the file folder:

```
<PAMDataset>
  <PAMRasterBand band="1">
    <Metadata>
      <MDI key="STATISTICS_MAXIMUM">835</MDI>
      <MDI key="STATISTICS_MEAN">37.024768992184</MDI>
      <MDI key="STATISTICS_MINIMUM">2</MDI>
      <MDI key="STATISTICS_STDDEV">71.52569727799</MDI>
    </Metadata>
  </PAMRasterBand>
</PAMDataset>
```

5) After this, delete the XML file created by gdalinfo.

6) Load the perigosidade_int.tif raster file in QGIS.

7) You don't need to do anything inside QGIS, just remove the raster from QGIS TOC.

8) If you go to the folder where the raster is stored, QGIS created a new XML file with:

```
<PAMDataset>
  <PAMRasterBand band="1">
    <Histograms>
      <HistItem>
        <HistMin>1.500685871056241</HistMin>
        <HistMax>730.4993141289438</HistMax>
        <BucketCount>729</BucketCount>
        <IncludeOutOfRange>0</IncludeOutOfRange>
        <Approximate>1</Approximate>
        <HistCounts>(...)</HistCounts>
      </HistItem>
    </Histograms>
    <Metadata>
      <MDI key="STATISTICS_MAXIMUM">730</MDI>
      <MDI key="STATISTICS_MEAN">34.375557670573</MDI>
      <MDI key="STATISTICS_MINIMUM">2</MDI>
      <MDI key="STATISTICS_STDDEV">67.853826109685</MDI>
    </Metadata>
  </PAMRasterBand>
</PAMDataset>
```

9) So, QGIS marks the **STATISTICS_MAXIMUM** as **730**.

10) Even defining in Settings -> Options -> Rendering -> Rasters -> Limits (minimum/maximum): Minimum/Maximum

QGIS always uses 730, because the Accuracy is, by default, "Estimate (faster)", and this option is not configurable at Options, as said in the related tickets.

But what is really serious here, is that as soon as the XML file is created, Processing tools use these statistics for the calculations.

11) For instance, if run r.quantile with "generate recode rules" flag, the output is:

```
2.000000:6.000000:1
6.000000:8.000000:2
8.000000:12.000000:3
12.000000:20.000000:4
20.000000:730.000000:5
```

12) Classifying the raster based on these rules with r.recode leaves pixels with value 835 as NULL.

And every subsequent process gives errors.

The question is.. how I never had realized this problem? The fact is that QGIS only generates the XML file when the raster layer is removed from QGIS project. And so, if Processing modules are runned using the layers loaded in TOC, there are no problems.

If these modules are runned from the python console (processing.runalg / processing.run), from a plugin or using raster layers that are not loaded in QGIS, the problem show up.

This was tested in QGIS 2.18.22 and 3.2.1 (OSGeo4W64).

It seems a really serious problem to me.

Associated revisions

Revision 70d4a276 - 2018-10-19 10:45 AM - Matthias Kuhn

Do not persist estimated GDAL metadata

GDAL saves metadata like min and max values into a .aux.xml sidecar file next to raster files.
It does this always, even when the calculated values are estimated. In subsequent runs of GDAL processing tools
it will use these values as if they were reliable.

This patch takes care of deleting newly written .aux.xml files if there is a risk that they include estimated data.

Fix #19517 https://issues.qgis.org/issues/19517

Revision 87fddaee - 2018-10-25 02:02 PM - Matthias Kuhn

Do not persist estimated GDAL metadata

GDAL saves metadata like min and max values into a .aux.xml sidecar file next to raster files.
It does this always, even when the calculated values are estimated. In subsequent runs of GDAL processing tools
it will use these values as if they were reliable.

This patch takes care of deleting newly written .aux.xml files if there is a risk that they include estimated data.

Fix #19517 https://issues.qgis.org/issues/19517

History

#1 - 2018-08-05 08:11 PM - Giovanni Manghi

- Crashes QGIS or corrupts data changed from No to Yes

I'm tagging this as "corrupts" data, because wrong results are never acceptable.

#2 - 2018-08-10 09:50 PM - Jürgen Fischer

- Description updated

#3 - 2018-09-14 11:25 AM - Pedro Venâncio

This seems a really tricky issue and not trivial to solve.

But maybe in the meanwhile, it could be better to change the default Accuracy option to "Actual (slower)". It can slower QGIS a litle bit with huge rasters, but prevent issues with geoprocessing tools and wrong results.

#4 - 2018-10-15 12:48 AM - Giovanni Manghi

- Affected QGIS version changed from 2.18.21 to 3.2.1

#5 - 2018-10-18 02:46 PM - Matthias Kuhn

- Assignee set to Matthias Kuhn

- Affected QGIS version changed from 3.2.1 to 3.3(master)

It looks like GDAL writes this file when GDALClose is called.
I am considering to delete it right after if two conditions are met: the file was not there before and the statistics are estimated.

A real fix would probably involve GDAL (I guess it shouldn't permanently write estimated data at all?) but this should at least help for a bandaid fix.

#6 - 2018-10-18 03:30 PM - Matthias Kuhn

Pull request pending https://github.com/qgis/QGIS/pull/8230

#7 - 2018-10-25 10:20 AM - Anonymous

- % Done changed from 0 to 100

- Status changed from Open to Closed

Applied in changeset commit:qgis|70d4a276da8eaa4c5b3e84c467b95a3611b4b9c7.