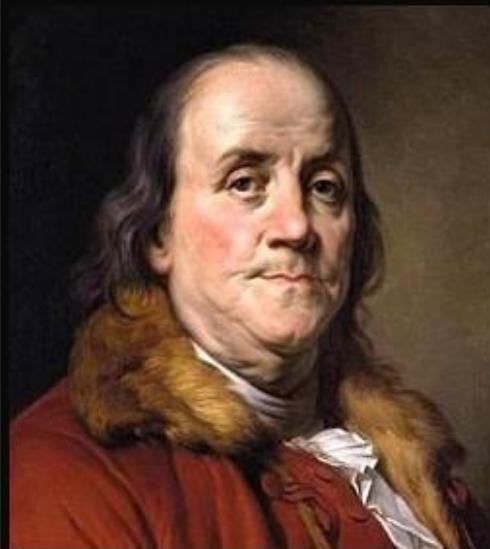


A Survey of QGIS Methods for Detecting Data Quality Errors



Our new Constitution is now established, and has an appearance that promises permanency; but in this world nothing can be said to be certain, except death and taxes.
and map data errors!

(Benjamin Franklin)

izquotes.com

A Survey of QGIS Methods for Detecting Data Quality Errors

1

- The opening question:
- *What are you looking for?*

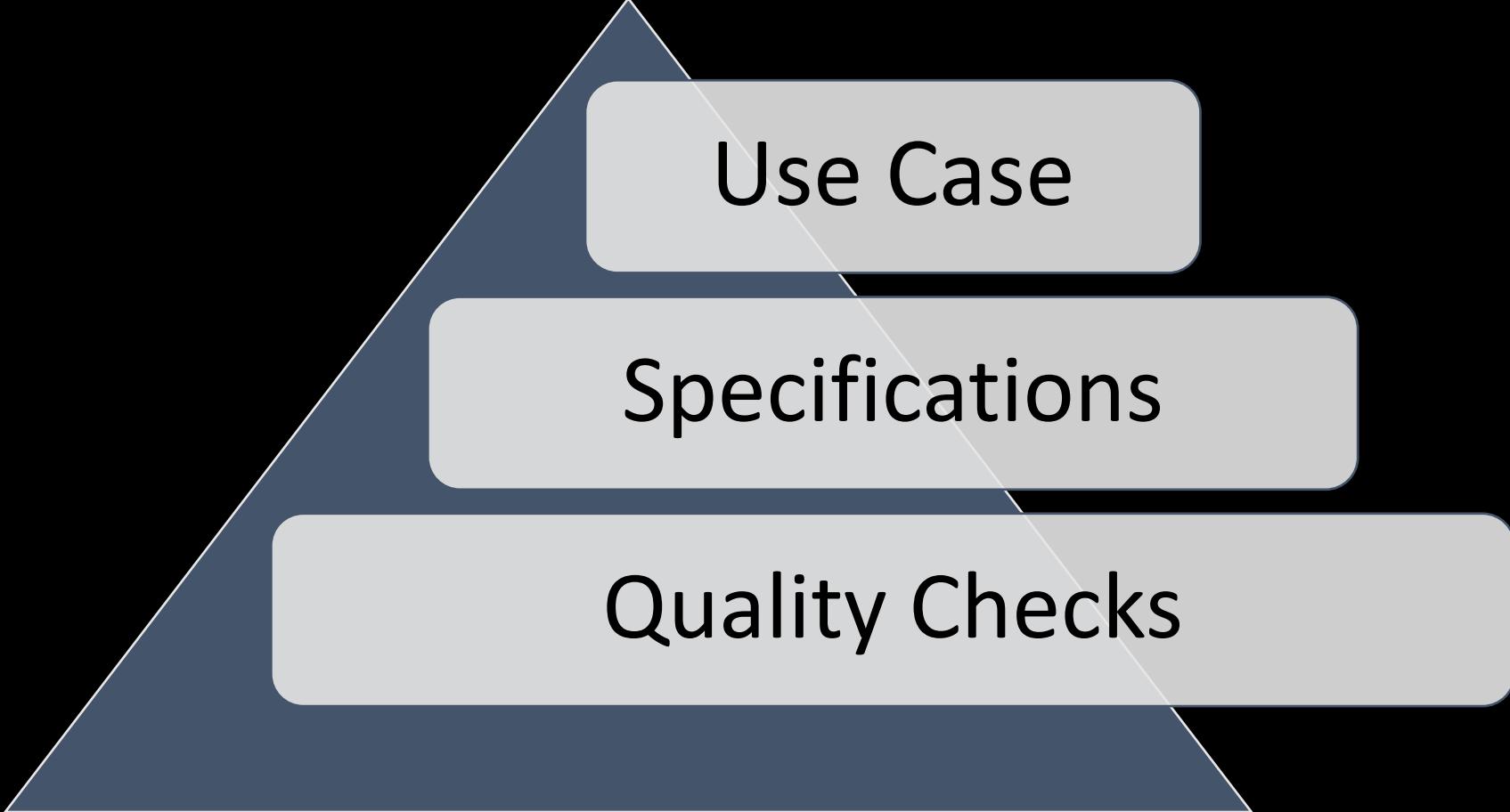
2

- Examples:
- *Specs and Checks*

3

- The closing question:
- *What is your framework?*

1. What are you looking for?

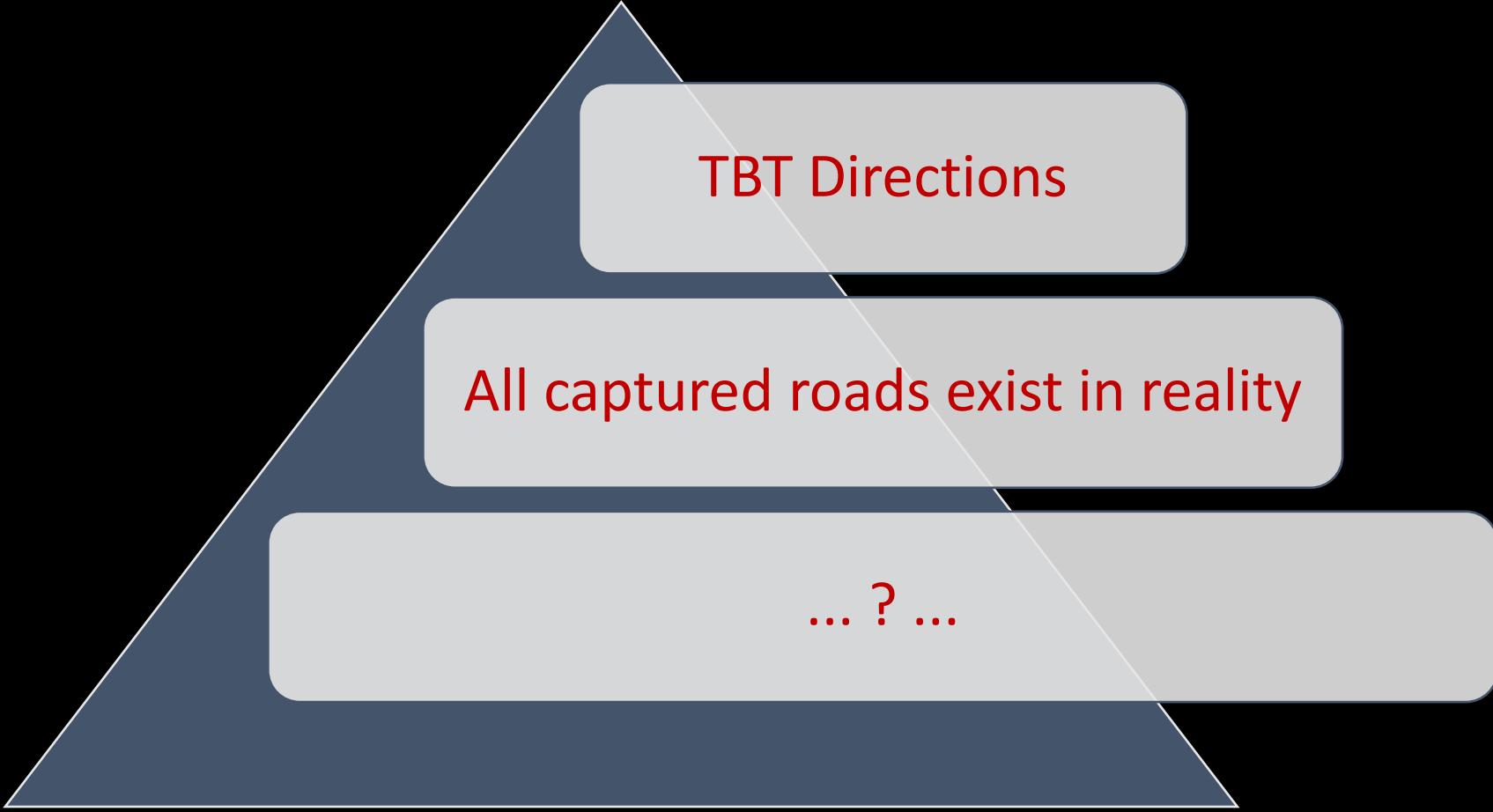


Use Case

Specifications

Quality Checks

1. What are you looking for?

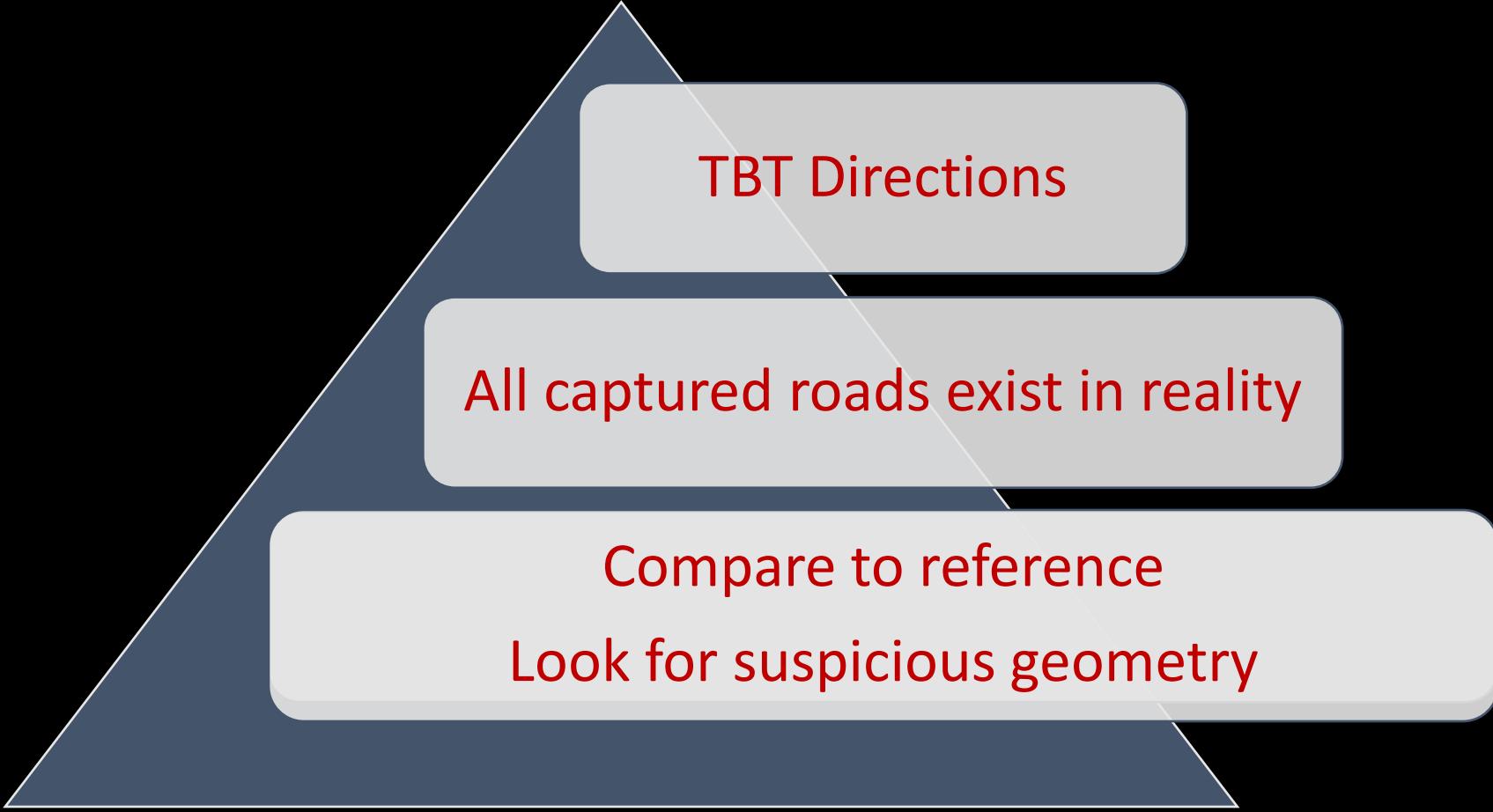


TBT Directions

All captured roads exist in reality

... ? ...

1. What are you looking for?



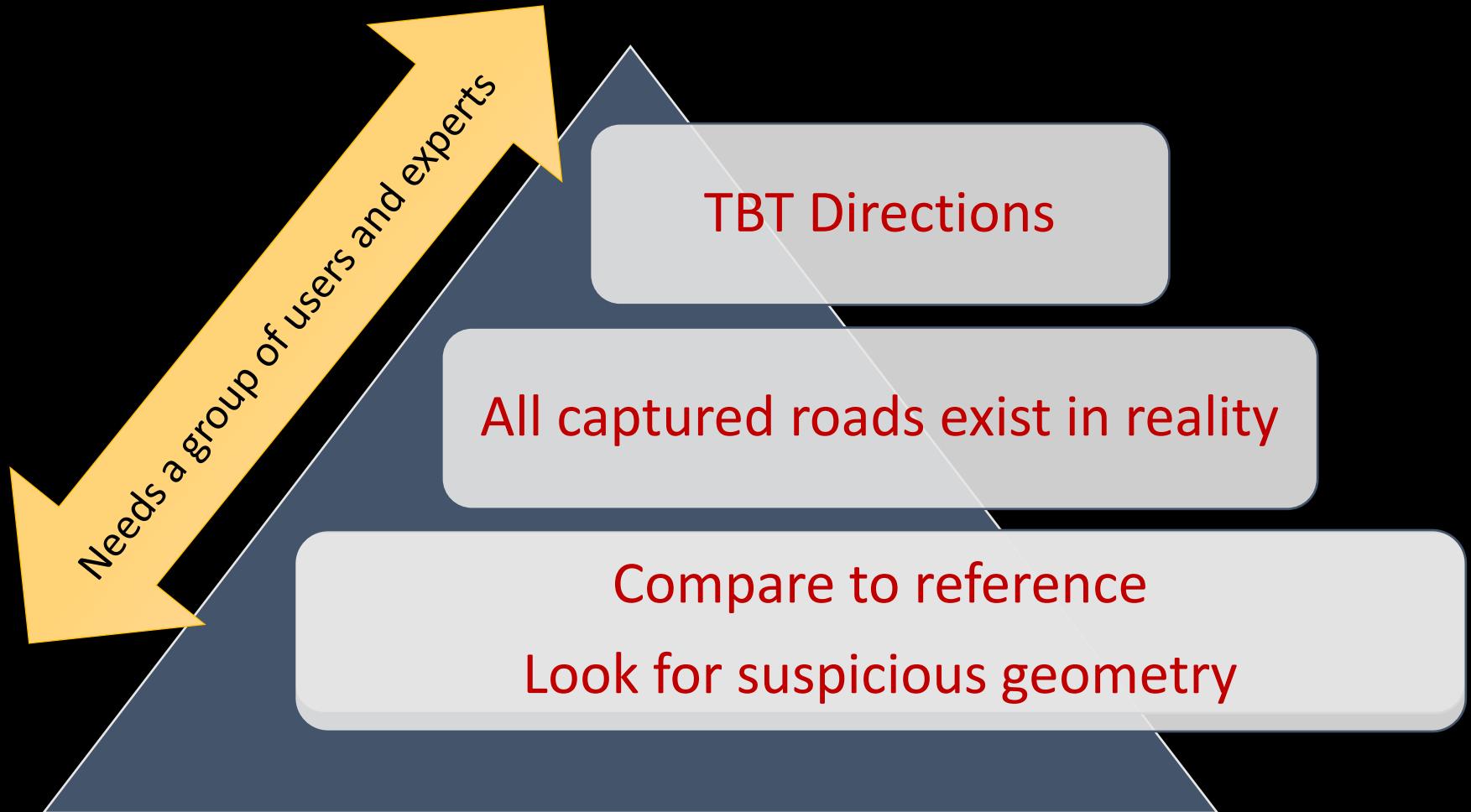
TBT Directions

All captured roads exist in reality

Compare to reference

Look for suspicious geometry

1. What are you looking for?





SCIENTIFIC
AMERICAN™

SUBSCRIBE



TECH

5 Most Embarrassing Software Bugs in History

Most software today arrives full of small bugs. But big glitches have lost whole spacecraft or could send tourists driving into the ocean

Apple Maps gives us directions to nowhere (2012): In 2012, when Apple decided to ditch Google Maps in favor of its own map app, it was a major blow to users. After all, Google Maps had been the go-to map app for years. But Apple's new map app was plagued by bugs, including one that caused it to give directions to nowhere. This was a major embarrassment for Apple, and it helped to fuel the company's decision to bring back Google Maps in 2013.

A Survey of QGIS Methods for Detecting Data Quality Errors

1

- The opening question:
- *What are you looking for?*

2

- Examples:
- *Specs and Checks*

3

- The closing question:
- *What is your framework?*

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

A. City population values are realistic.

CHECK: For a City, the population should not be less than 3000.

A. City population values are realistic.

CHECK: For a City, the population should not be less than 3000.

For a FEATURE,
the ATTRIBUTE
[must/should]
be CONSTRAINT.

A. City population values are realistic.

CHECK: For a City, the population should not be less than 3000.

For a FEATURE,
the ATTRIBUTE
[must/should]
be CONSTRAINT.

METHOD: Sort the
attribute table.

Attribute table - cities :: Features total: 606, filtered: 606, selected: 0				
	NAME	COUNTRY	POPULATION	CAPITAL
464	Salekhard	Russia	-99	N
474	Tiksi	Russia	-99	N
389	Wadi Halfa	Sudan	-99	N
354	Chingmei	Taiwan	-99	N
348	Kahemba	Zaire	-99	N
432	Godhavn	Greenland	1012	N
424	Churchill	Canada	1304	N
336	Schefferville	Canada	1997	N
416	Angmagssalik	Greenland	2618	N
470	Seward	US	2699	N
449	Nome	US	3500	N
494	Belmopan	Belize	4500	Y
435	Hammerfest	Norway	7208	N

A. City population values are realistic.

CHECK: For a City, the population should not be less than 3000.

For a FEATURE,
the ATTRIBUTE
[must/should]
be CONSTRAINT.

METHOD: Sort the
attribute table.

Attribute table - cities :: Features total: 606, filtered: 606, selected: 0				
	NAME	COUNTRY	POPULATION	CAPITAL
464	Salekhard	Russia	-99	N
474	Tiksi		-99	N
389	Wadi Halfa			
354	Chingmei			
348	Kahemba			
432	Godhavn			
424	Churchill			
336	Schefferville			
416	Angmagssalik			
470	Seward			
449	Nome			
494	Belmopan	Belize	4500	Y
435	Hammerfest	Norway	7208	N

Basic data
inspection =

plan for it as the
first step

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are from a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

B. Road attributes values are from a defined list.

CHECK: For a Road, the MTFCC value must be one of the following: S1200, S1400, ...

*Precise and
unambiguous =
Everyone will get
the same results.*

B. Road attributes values are from a defined list.

CHECK: For a Road, the MTFCC value must be one of the following: S1200, S1400, ...

METHOD 1:

Select by Expression

METHOD 2:

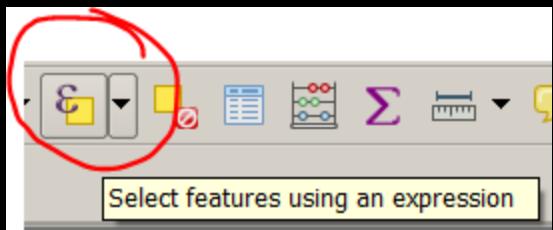
Custom Expression

METHOD 3:

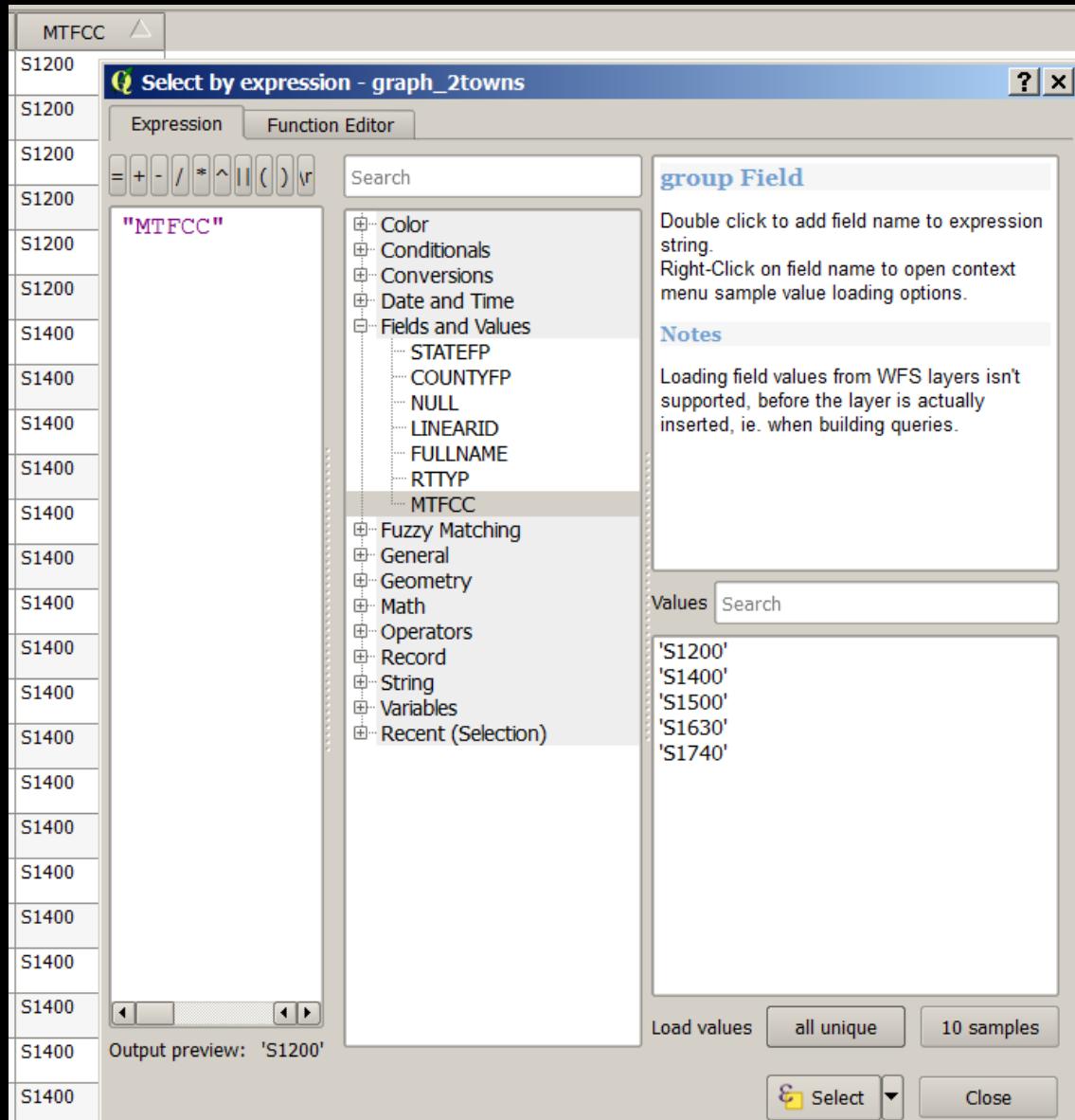
Plugin: GroupStats

*Precise and
unambiguous =
Everyone will get
the same results.*

METHOD 1



- Select By Expression
- Fields and Values
- Load values: all unique
- ... finish expression with any invalid values



METHOD 2

The screenshot shows the QGIS Plugins Manager interface. At the top, a menu bar has 'Plugins' selected. Below it is a toolbar with icons for 'Manage and Install Plugins...', 'Python Console' (with keyboard shortcut 'Ctrl+Alt+P'), and 'NNJoin'. The main window title is 'Plugins | All (567)'. On the left, a sidebar lists filter categories: 'All' (selected), 'Installed', 'Not installed', 'Upgradeable', 'Invalid', and 'Settings'. A search bar contains the text 'group'. The main pane displays a list of plugins, with 'Group Stats' highlighted. To the right, a detailed view of the 'Group Stats' plugin is shown, including its name, rating, category, tags, author, and installed status.

Plugins | All (567)

All Search group

Installed
Not installed
Upgradeable
Invalid
Settings

Clusterpy - Spatially constraine
Dissolve with stats
Floodrisk
Group Stats
ImportLayersFromProject
Layer arranger
Layer Metadata Dock
Loop Visible Layers
NTv2 Datum Transformations
PointsToPaths
prepair
Slicer

Group Stats

Stats and analysis for vector layers data

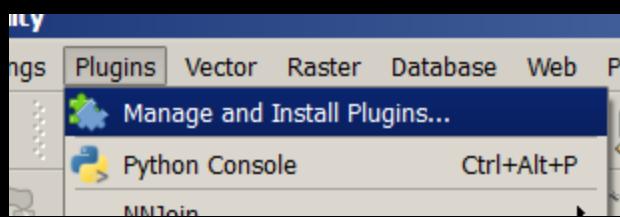
★★★★★ 51 rating vote(s), 69763 downloads

Category: Vector
Tags: stats,analysis,pivot table
More info: [homepage](#) [tracker](#)

Author: [Rajmund Szostok](#)

Installed version: 2.0.30 (in C:\Users\andresmi\qgis2\python\plugins\GroupStats)
Available version: 2.0.30 (in QGIS Official Plugin Repository)

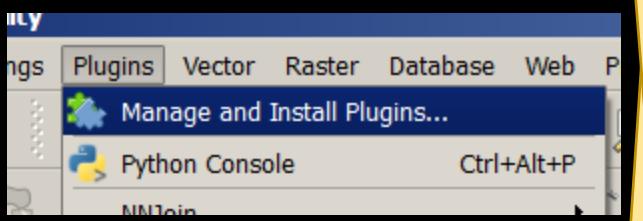
METHOD 2



Also searchable at
<http://plugins.qgis.org/plugins>

A screenshot of the QGIS Plugins manager. The title bar says 'Plugins | All (567)'. On the left, there's a sidebar with filters: 'All' (selected), 'Installed', 'Not installed', 'Upgradeable', 'Invalid', and 'Settings'. A search bar at the top right contains the text 'group'. Below the search bar is a list of plugins. The 'Group Stats' plugin is selected, indicated by a blue highlight and a small 'X' icon. Other listed plugins include 'Clusterpy - Spatially constraine', 'Dissolve with stats', 'Floodrisk', 'ImportLayersFromProject', 'Layer arranger', 'Layer Metadata Dock', 'Loop Visible Layers', 'NTv2 Datum Transformations', 'PointsToPaths', 'prepair', and 'Slicer'. To the right of the search results, there's a detailed view for the 'Group Stats' plugin. It has a title 'Group Stats' with a question mark icon. Below it is a bold heading 'Stats and analysis for vector layers data'. There's a 5-star rating icon followed by '51 rating vote(s), 69763 downloads'. Underneath, it lists 'Category: Vector', 'Tags: stats,analysis,pivot table', and 'More info: [homepage](#) [tracker](#)'. The author is listed as 'Author: [Rajmund Szostok](#)'. At the bottom, it shows the 'Installed version: 2.0.30 (in C:\Users\andresmi\qgis2\python\plugins\GroupStats)' and 'Available version: 2.0.30 (in QGIS Official Plugin Repository)'.

METHOD 2

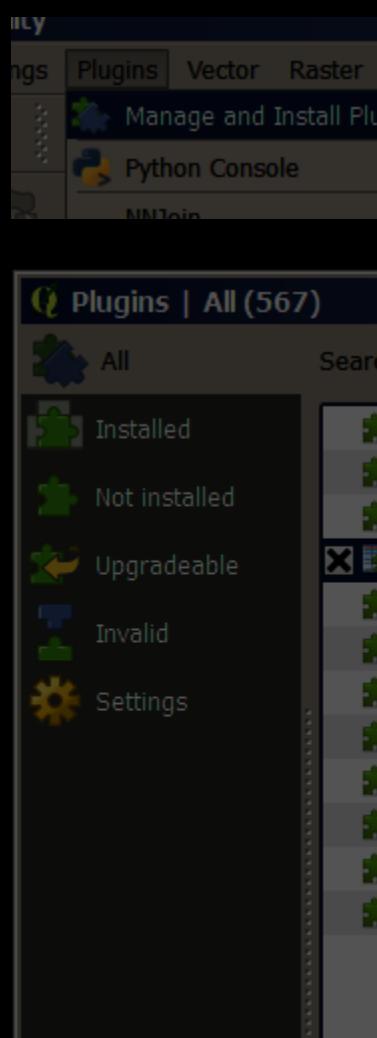


Or build your own plugin:

http://www.qgistutorials.com/en/docs/building_a_python_plugin.html

A screenshot of the QGIS Plugins manager. The title bar says 'Plugins | All (567)'. On the left, there is a sidebar with categories: All, Installed, Not installed, Upgradeable, Invalid, and Settings. The 'All' category is selected. A search bar above the list contains the text 'group'. The main list shows various plugins, with 'Group Stats' selected and highlighted with a dark blue background. To the right of the list, there is a detailed view of the 'Group Stats' plugin. It has a title 'Group Stats', a subtitle 'Stats and analysis for vector layers data', and a rating of 5 stars with 51 votes and 69763 downloads. It is categorized under 'Vector' and has tags for 'stats', 'analysis', and 'pivot table'. There are links for 'More info', 'homepage', and 'tracker'. The author is listed as 'Rajmund Szostok'. The installed version is 2.0.30 (in C:\Users\andresmi\qgis2\python\plugins\GroupStats). An available version of 2.0.30 is also mentioned.

METHOD 2



Group Stats

Data Features Window Help

Control panel

Layers: graph_2towns

Fields:

- average
- count
- max
- median
- min
- stand.dev.
- sum
- unique
- variance

Filter

Columns:

- count
- sum

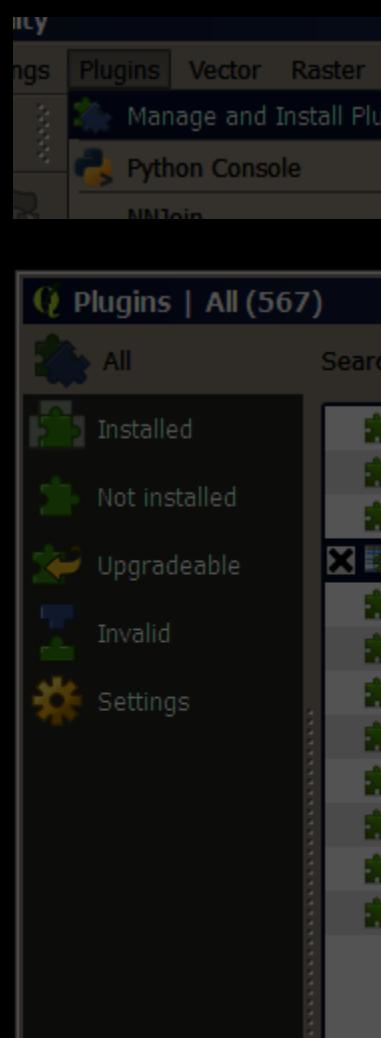
Rows:

Value use NULL values

MTFCC Length

This screenshot shows the "Group Stats" dialog box from QGIS. The main area displays a table with columns labeled 1, 2, and 3. Row 1 contains the header "Function" with values "count" and "sum". Row 2 contains the value "MTFCC". Rows 3 through 7 contain values for S1200, S1400, S1500, S1630, and S1740 respectively. The "Fields" panel on the right lists various statistical functions: average, count, max, median, min, stand.dev., sum, unique, and variance. The "Columns" panel shows "count" and "sum" selected. The "Rows" panel shows "MTFCC" and "Length". A "Filter" button is also present.

METHOD 2



Group Stats

Data Features Window Help

Control panel

1	2	3
Function	count	sum
2 MTFCC		
3 S1200	6	29307.1
4 S1400	191	156246
5 S1500	9	13108.5
6 S1630	1	25.9093
7 S1740	75	15876.2

Layers
graph_2towns

Fields

- average
- count
- max
- median
- min
- stand.dev.
- sum
- unique
- variance

Filter

Columns

Value use NULL values

Rows

Value use NULL values

MTFCC Length

A yellow sticky note is overlaid on the interface, containing handwritten text:

Like an EXCEL PivotTable

Count +
sum(length)
indicates scope

The screenshot shows a GIS application interface. On the left is a map of a river network, with a specific segment highlighted by a thick yellow line. To the right of the map is a window titled "Group Stats" containing a table of statistics. Below the map and the "Group Stats" window is a "Control panel" window.

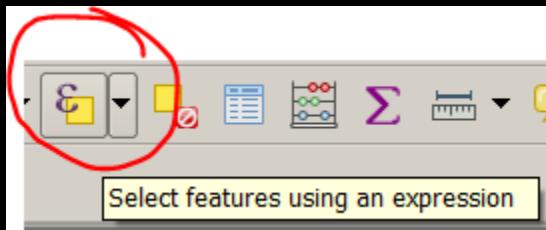
Group Stats Window:

Function	count	sum
MTFCC		
S1200	6	29307.1
S1400	191	156246
S1500	9	13108.5
S1630	1	25.9093
S1740	75	15876.2

Control Panel Window:

- Layers:** A yellow sticky note with handwritten text: "*← click a row then Features > show selected on map*".
- variance**: A section with a "Filter" button and a "Columns" section containing "count" and "sum".
- Rows**: A section with a "Value" checkbox labeled "use NULL values".
- MTFCC**: A section with a "Length" button.

METHOD 3



Select by expression - graph_2towns

Expression **Function Editor** (circled in red)

New file Load

NewFunction

- AttributeRange
- MTFCC_range

```
1 """  
2 Define new functions using @qgsfunction. feature and parent must always be the  
3 last args. Use args=-1 to pass a list of values as arguments  
4 """  
5  
6 from qgis.core import *  
7 from qgis.gui import *  
8  
9 @qgsfunction(args='auto', group='Custom')  
10 - def func(value1, feature, parent):  
11     return value1  
12
```

METHOD 3

Q Select by expression - graph_2towns

Expression Function Editor

New file Load

NewFunction AttributeRange MTFCC_range

```
1 """
2 Define new functions using @qgsfunction. feature and parent must always be the
3 last args. Use args=-1 to pass a list of values as arguments
4 """
5
6 from qgis.core import *
7 from qgis.gui import *
8
9 @qgsfunction(args='auto', group='Custom')
10 def MTFCCHasAllowedValue(value1, feature, parent):
11     """
12         Returns a boolean indicating whether the attribute value
13         is part of the defined range for the provided field.
14     """
15     allowed_values = ['S1100', 'S1200']
16     found_value = feature['MTFCC']
17     isAllowed = found_value in allowed_values
18     return isAllowed
```

METHOD 3

Select by expression - graph_2towns

Expression Function Editor

New file Load

NewFunction AttributeRange MTFCC_range

```
1 """  
2 Define new functions using @qgsfunction  
3 last args. Use args=-1 to pass a list of  
4 """  
5  
6 from qgis.core import *  
7 from qgis.gui import *  
8  
9 @qgsfunction(args='auto', group='Custom')  
10 def MTFCCHasAllowedValue(value1, feature, parent):  
11     """  
12         Returns a boolean indicating whether the attribute value  
13         is part of the defined range for the provided field.  
14     """  
15     allowed_values = ['S1100', 'S1200']  
16     found_value = feature['MTFCC']  
17     isAllowed = found_value in allowed_values  
18     return isAllowed
```

- Click Load
- Function remains available
- File is auto-saved in .qgis2/python/expressions
- File can be imported elsewhere

METHOD 3

Select by expression - graph_2towns

Expression Function Editor

= + - / * ^ || () '\n'

MTFCCHasAllowedValue("MTFCC") = False

Search

- + Color
- + Conditionals
- + Conversions
- + Custom
 - FieldHasAllowedValue
 - MTFCCHasAllowedValue
- + Date and Time
- + Fields and Values
 - STATEFP
 - COUNTYFP
 - NULL
 - LINEARID
 - FULLNAME
 - RTTYP
 - MTFCC
 - LINE_LEN
- + Fuzzy Matching
- + General
- + Geometry
- + Math
- + Operators
- + Record
- + String
- + Variables
- + Recent (Selection)

MTFCCHasAllowedValue function

Returns a boolean indicating whether the attribute value is part of the defined range for the provided field.

METHOD 3

Really powerful!

Resources:

- [http://www.qgistutorials.com/en/docs/custom python functions.html](http://www.qgistutorials.com/en/docs/custom_python_functions.html)
- <https://nathanw.net/2012/11/10/user-defined-expression-functions-for-qgis/>
- <https://nathanw.net/2015/01/19/function-editor-for-qgis-expressions/>

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are from a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

2. Specs and Checks

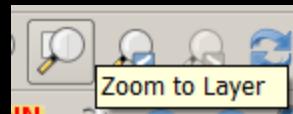
- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

C. Roads are captured within 10m of their locations in reality.

CHECK: For a Road, all nodes and vertices must be located inside the bounding box defined by (x,y),(x,y)

METHOD 1:

Zoom to Layer

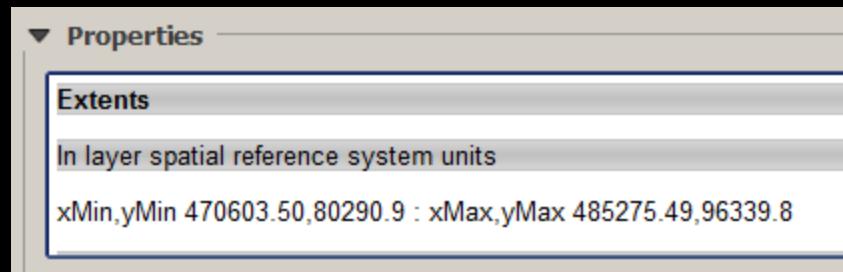


METHOD 2:

Layer Properties >

Metadata >

Extents



C. Roads are captured within 10m of their locations in reality.

CHECK: For a Road, all nodes and vertices must be located inside the bounding box defined by (x,y),(x,y)

METHOD 1:

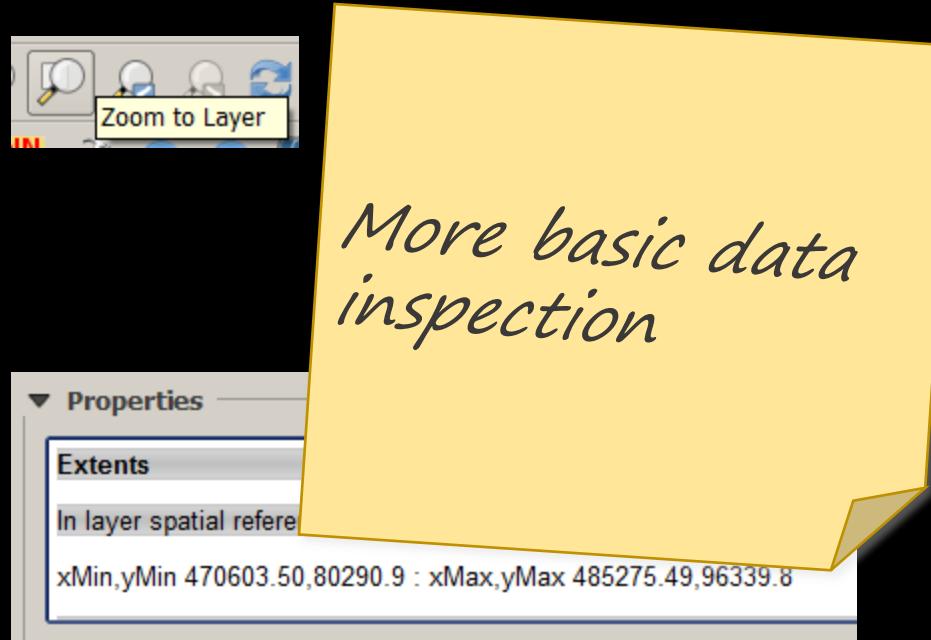
Zoom to Layer

METHOD 2:

Layer Properties >

Metadata >

Extents



C. Roads are captured within 10m of their locations in reality.

CHECK: For a Road, the length should not be greater than 15 kilometers.

METHOD: Field Calculator

The screenshot shows the QGIS interface with the Field Calculator open. The Field Calculator window has the following settings:

- Only update 0 selected features
- Create a new field
- Create virtual field
- Output field name: LINE_LEN
- Output field type: Decimal number (real)
- Output field length: 10
- Precision: 1

The Expression field contains the expression `len`. The function editor dropdown shows the following results:

- Geometry: \$length
- String: length
- Variables: project_filename

On the right side of the interface, there is a table titled "LINE_LEN" with the following data:

STATEFP	LINE_LEN
	13852.3
	8104.5
	7556.6
	6015.2
	5919.6
	5396.9
	5164.1
	4811.0
	4432.1
	4193.9
	3891.2
	3726.3
	3571.2
	3570.7

Below the table, a map view shows a network of roads. One specific road is highlighted in yellow, corresponding to the feature with the longest length in the table.

C. Roads are captured within 10m of their locations in reality.

CHECK: For a Road, the length should not be greater than 15 kilometers.

METHOD: Field Calculator

The screenshot shows the QGIS Field Calculator interface. The 'Create a new field' checkbox is selected, and the output field name is set to 'LINE_LEN'. The output field type is 'Decimal number (real)' with a length of 10 and precision of 1. The expression bar contains the variable '\$length'. A note is overlaid on the right side of the calculator window, reading: 'Remember to "Update Existing Field" after editing map data'.

Remember to
"Update Existing
Field" after
editing map data

Line ID	Length (km)
1	3891.2
2	3726.3
3	3571.2
4	3570.7

C. Roads are captured within 10m of their locations in reality.

CHECK: For a Road, the length should not be greater than 15 kilometers.

METHOD: Field Calculator

The screenshot shows the QGIS Field Calculator interface. The 'Create a new field' checkbox is selected, and the output field name is set to 'LINE_LEN'. The output field type is 'Decimal number (real)' with a length of 10 and precision of 1. The expression bar contains the variable '\$length'. A large yellow callout bubble points from the text 'Resource:' towards the URL provided in the note.

Resource:

http://docs.qgis.org/2.8/en/docs/user_manual/working_with_vector/field_calculator.html

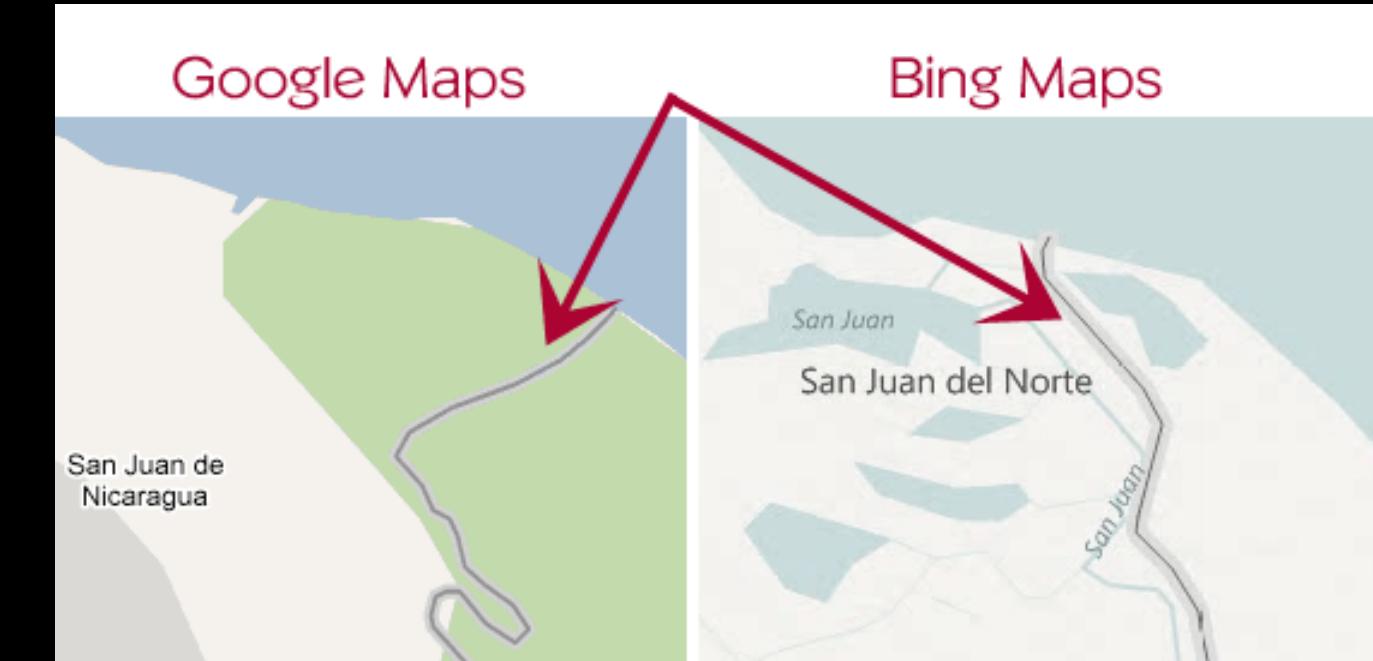
Feature ID	Length (m)
1	3891.2
2	3726.3
3	3571.2
4	3570.7

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.



Nicaraguan troops crossed the border, took down a Costa Rican flag and raised their own flag.

The Organization of American States and UN Security Council were asked to mediate.

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

D. Captured roads exist in reality.

CHECK: For a Road, the length should not be less than 5 meters.

METHOD 1: Calculated Field > Sort by Length

Attribute table - graph_2towns :: Features total: 282, filtered: 282, selected: 0							
	STATEFP	COUNTYFP	LINEARID	FULLNAME	RTTYP	MTFCC	LINE_LEN
60	50	027	110373961677	Schoolhouse Ln	M	S1400	0.5
107	50	027	110373986727	NULL	NULL	S1740	2.9
137	50	027	110373984542	NULL	NULL	S1400	8.3
251	50	027	110373955431	Park St	M	S1400	17.2
194	50	027	110373961566	Marsh Cross Rd	M	S1400	17.6
179	50	027	110373962539	Lovejoy Brook...	M	S1400	18.0
215	50	027	110373984110	NULL	NULL	S1400	19.6

D. Captured roads exist in reality.

CHECK: For a Road, the length should not be less than 5 meters.

METHOD 1: Calculating

Length

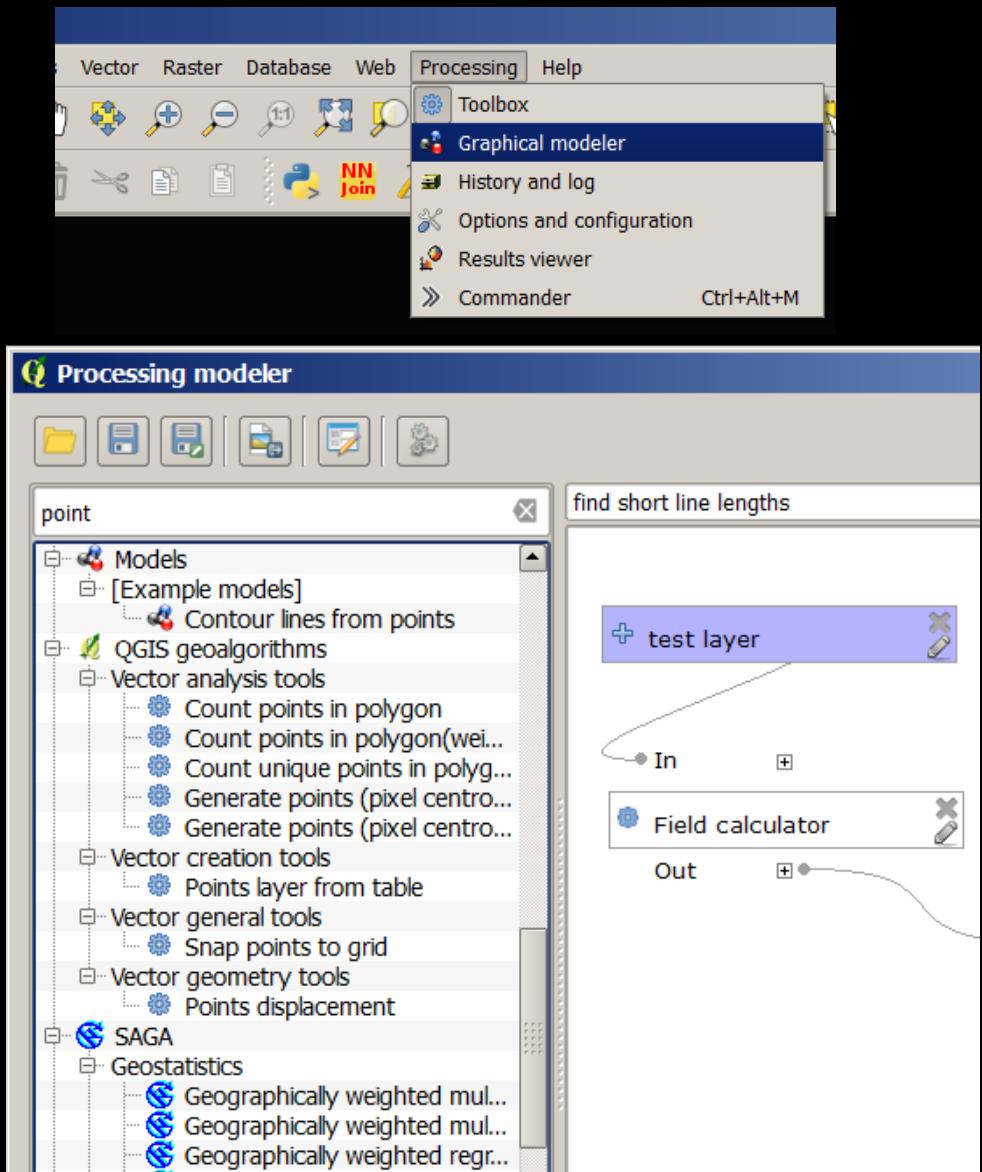
Not scalable

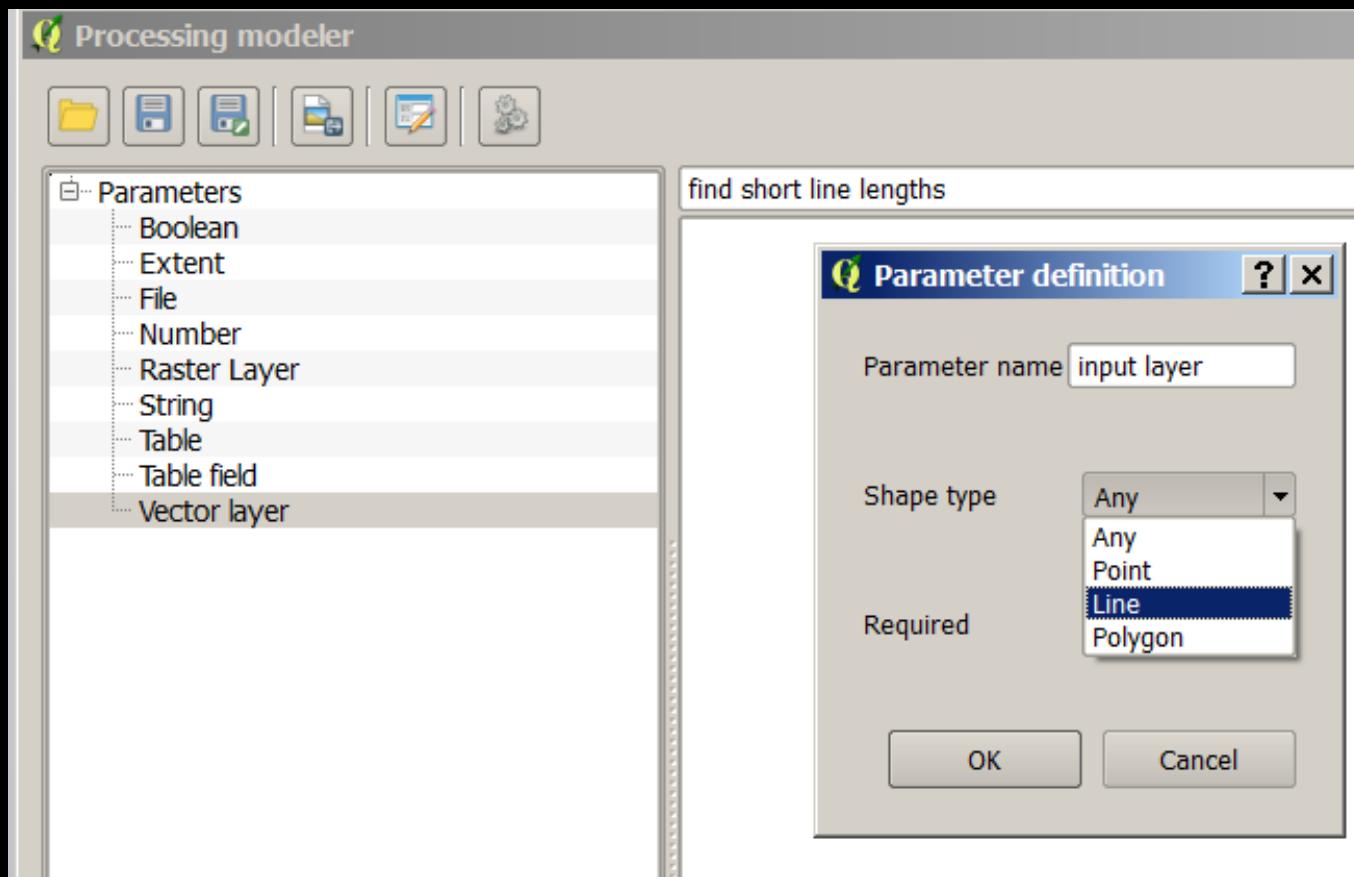
	STATEFP	COUNTYFP			TFCC	LINE_LEN	
60	50	027	110373984542	NULL	S1740	0.5	
107	50	027	110373984542	NULL	S1740	2.9	
137	50	027	110373984542	Park St	S1400	8.3	
251	50	027	110373955431	Marsh Cross Rd	S1400	17.2	
194	50	027	110373961566	Lovejoy Brook...	S1400	17.6	
179	50	027	110373962539	NULL	S1400	18.0	
215	50	027	110373984110	NULL	S1400	19.6	

D. Captured roads exist in reality.

METHOD 2: Model

- Create multi-step workflow
- Graphical interface
- Powerful processing tools
- Similar to ArcMap ModelBuilder FME Workbench





Q Processing modeler



Parameters

- Boolean
- Extent
- File
- Number
- Raster Layer
- String
- Table
- Table field
- Vector layer

find short line lengths

+ input layer 

Q Processing modeler



file

find short line lengths

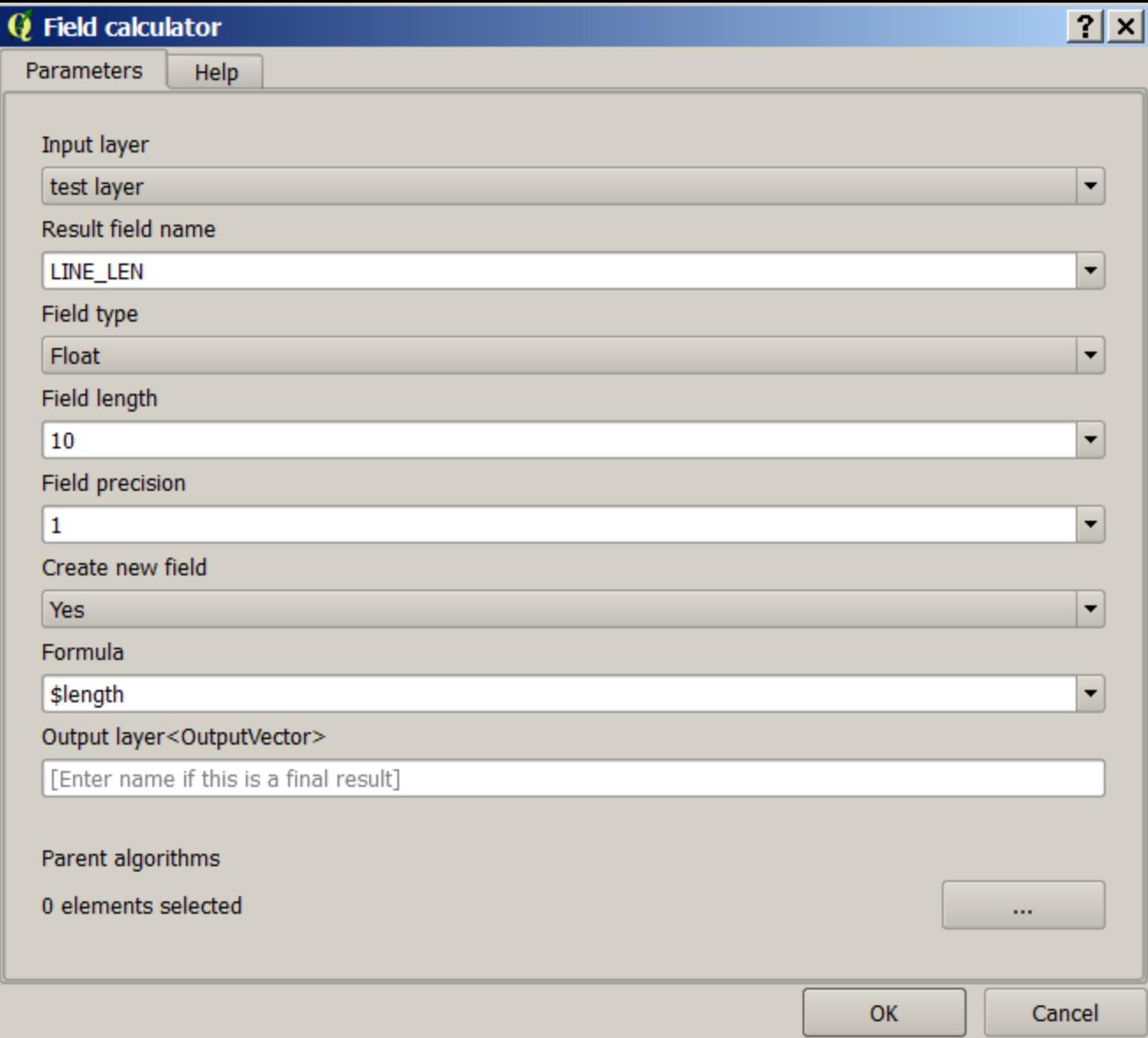
quality

- QGIS geoalgorithms
 - Vector table tools
 - Add autoincremental field
 - Add field to attributes table
 - Advanced Python field calculator
 - Basic statistics for numeric fields
 - Basic statistics for text fields
 - Create equivalent numerical field
 - Field calculator
- SAGA
 - Grid - Calculus
 - Random field
 - Table - Tools
 - Enumerate table field

+ test layer

+ length field

+ max length



Q Extract by attribute



Parameters

Help

Input Layer

Output layer from algorithm 0(Field calculator)

Selection attribute

length field

Comparison

<

Value

max length

Output<OutputVector>

output

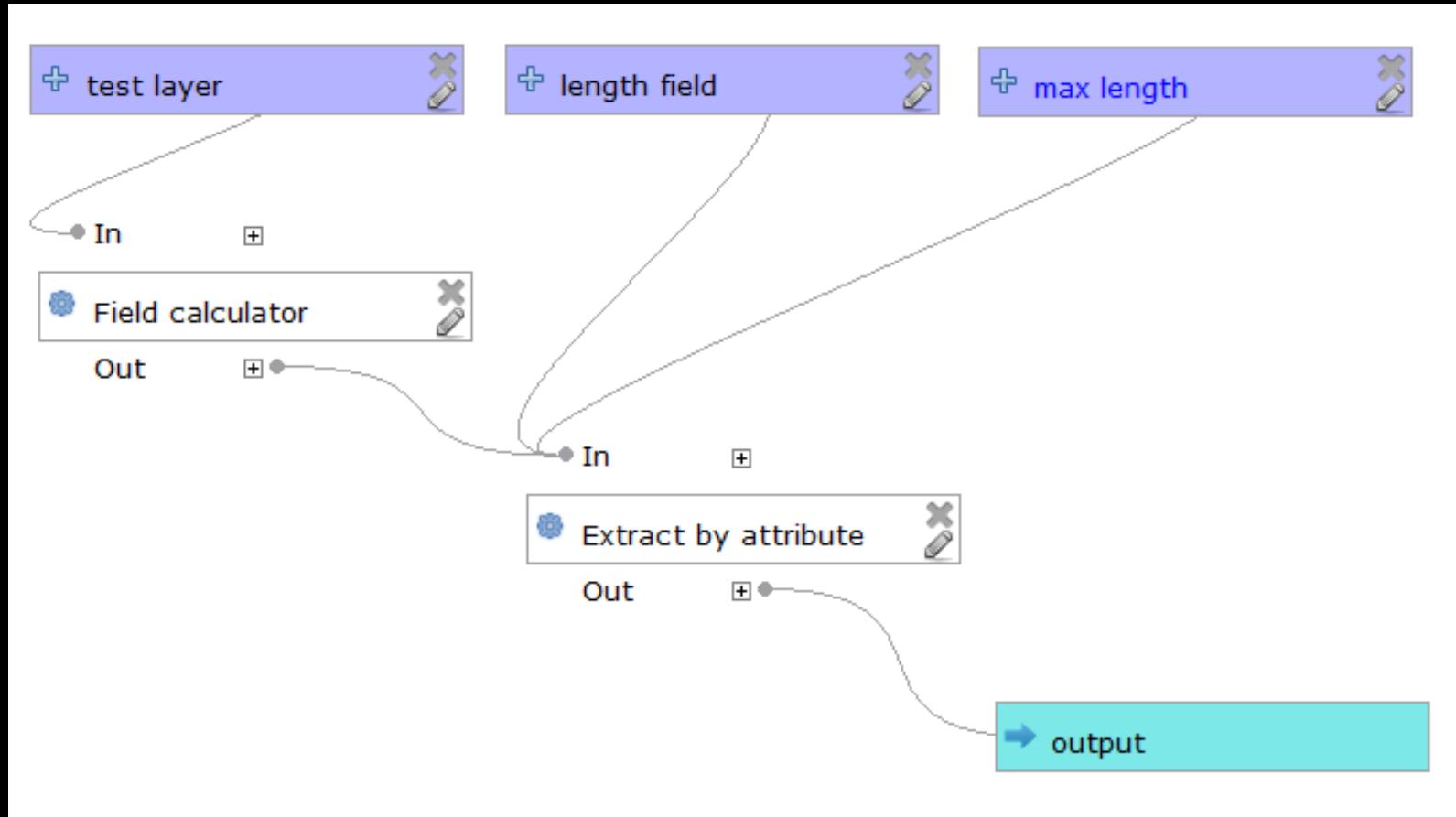
Parent algorithms

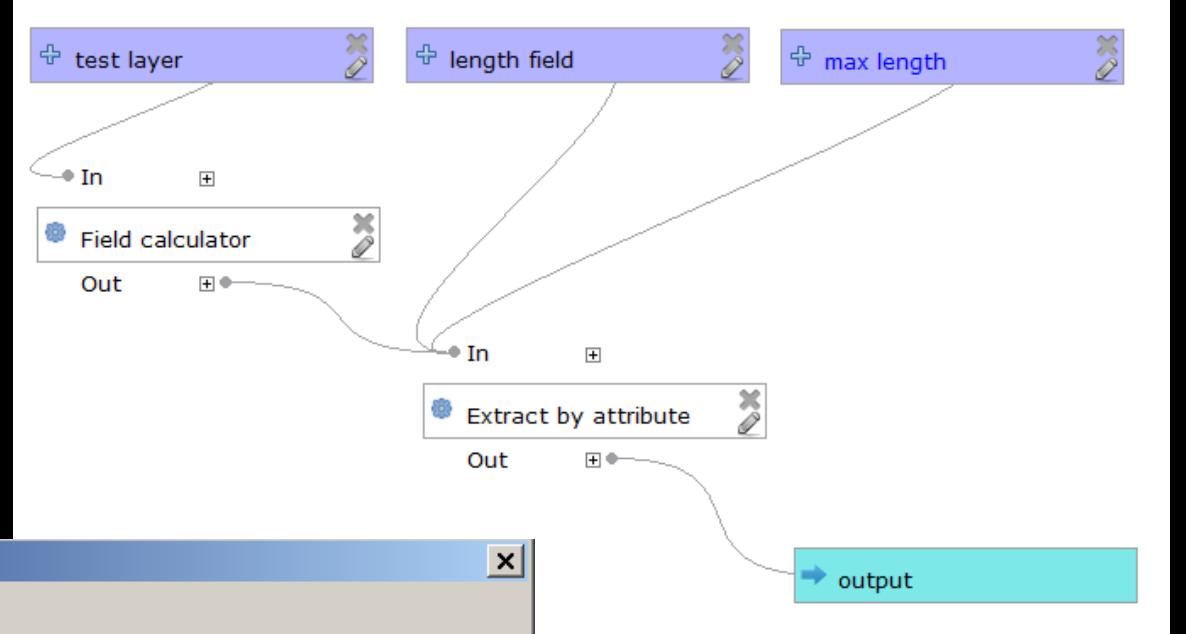
0 elements selected

...

OK

Cancel





find short line lengths

Parameters Log Help

test layer
graph_2towns [EPSG:102345]

max length
5

length field
LINE_LEN

output
s/andresmi/OneDrive - TomTom/analysis-projects/vcgi-quality/short_line_lengths.shp

Open output file after running algorithm

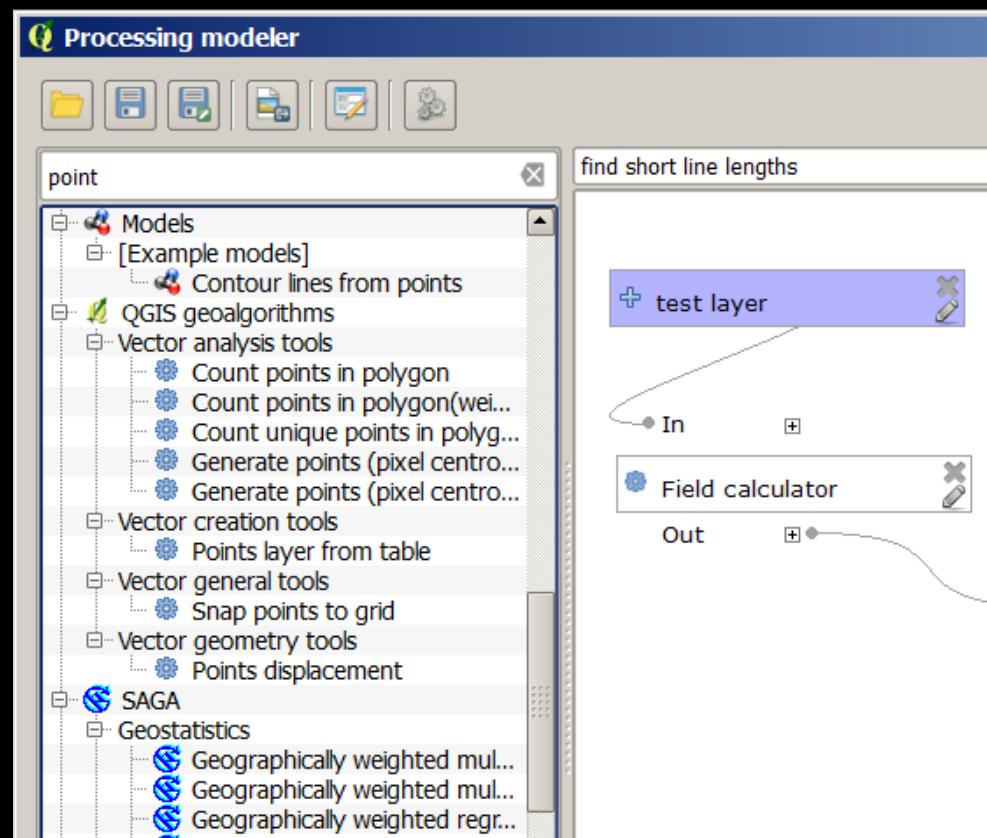
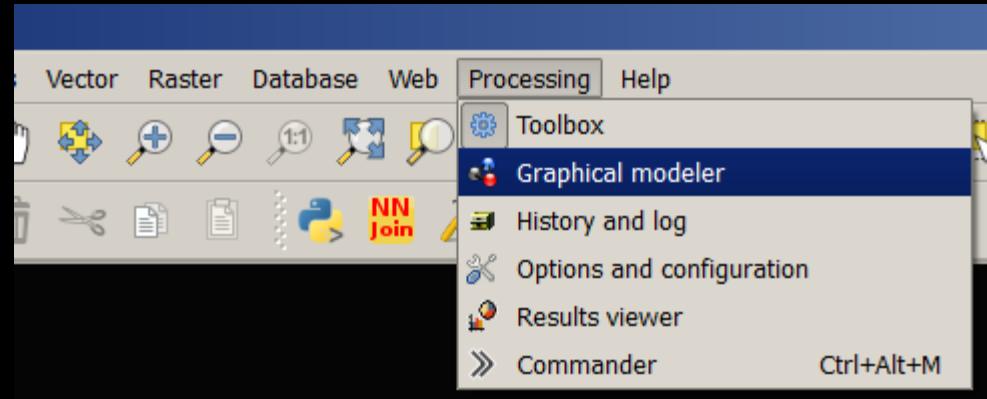
0%

Run Close Cancel

This panel displays the parameters for the "find short line lengths" algorithm. It includes fields for the input layer (graph_2towns), maximum length (5), length field (LINE_LEN), and output file (short_line_lengths.shp). A checkbox is present to open the output file after execution. At the bottom, there are buttons for Run, Close, and Cancel, along with a progress bar indicating 0% completion.

Models

- Repeatability
- Batch processing
- Data integrity
- Traceability
- Documentation
- Sharing
- Extendable in python



D. Captured roads exist in reality.

CHECK: For a Road, the entire feature must be captured within 10m of a feature in the designated reference dataset

METHOD: Buffer a reference dataset

D. Captured roads exist in reality.

CHECK: For a Road, the entire feature must be captured within 10m of a feature in the designated reference dataset

METHOD: Buffer

Positional
Accuracy

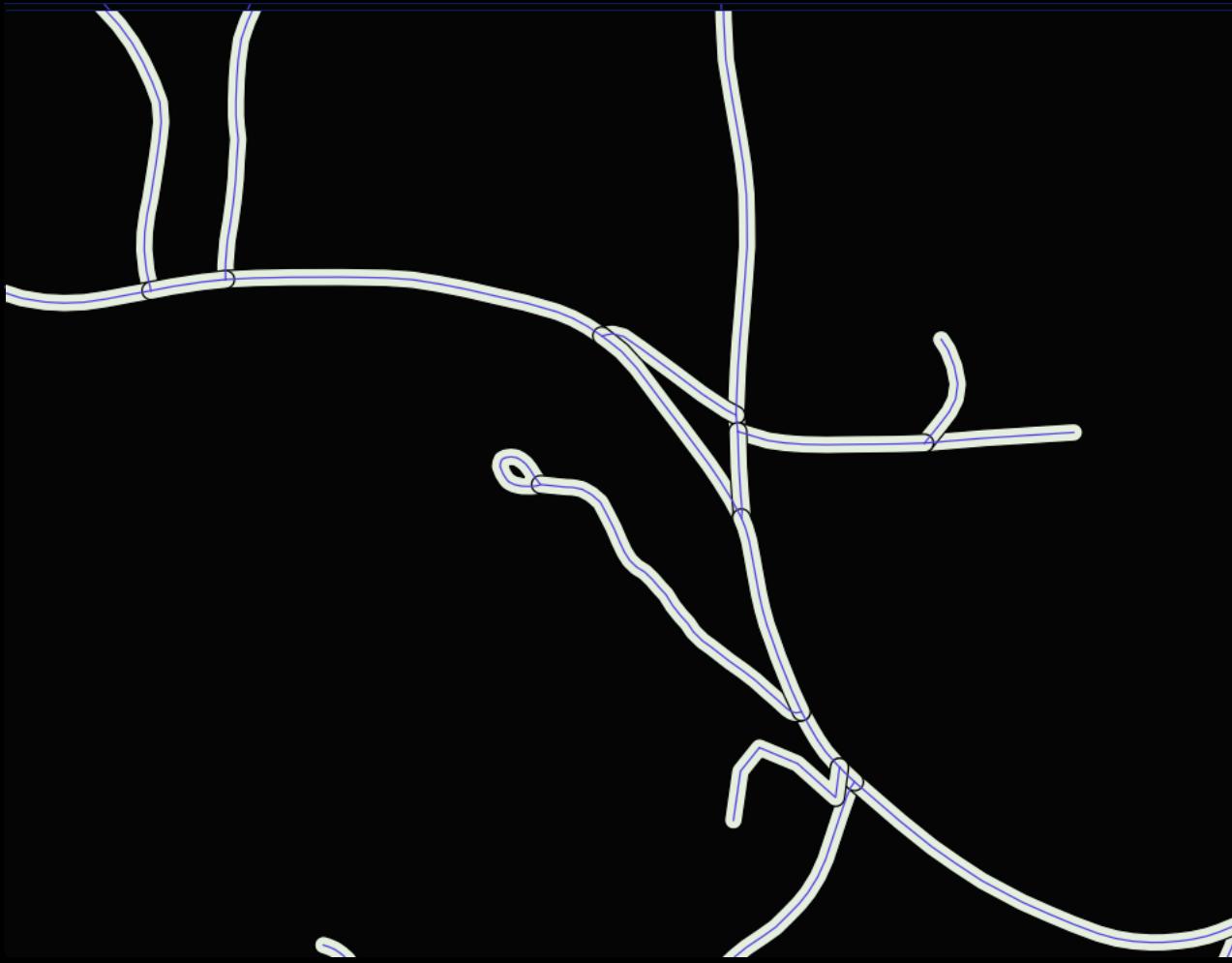
D. Captured roads exist in reality.

CHECK: For a Road, the entire feature must be captured within 10m of a feature in the designated reference dataset

METHOD: Buffer

<http://anitagraser.com/2013/12/21/osm-quality-assessment-with-qgis-network-length>



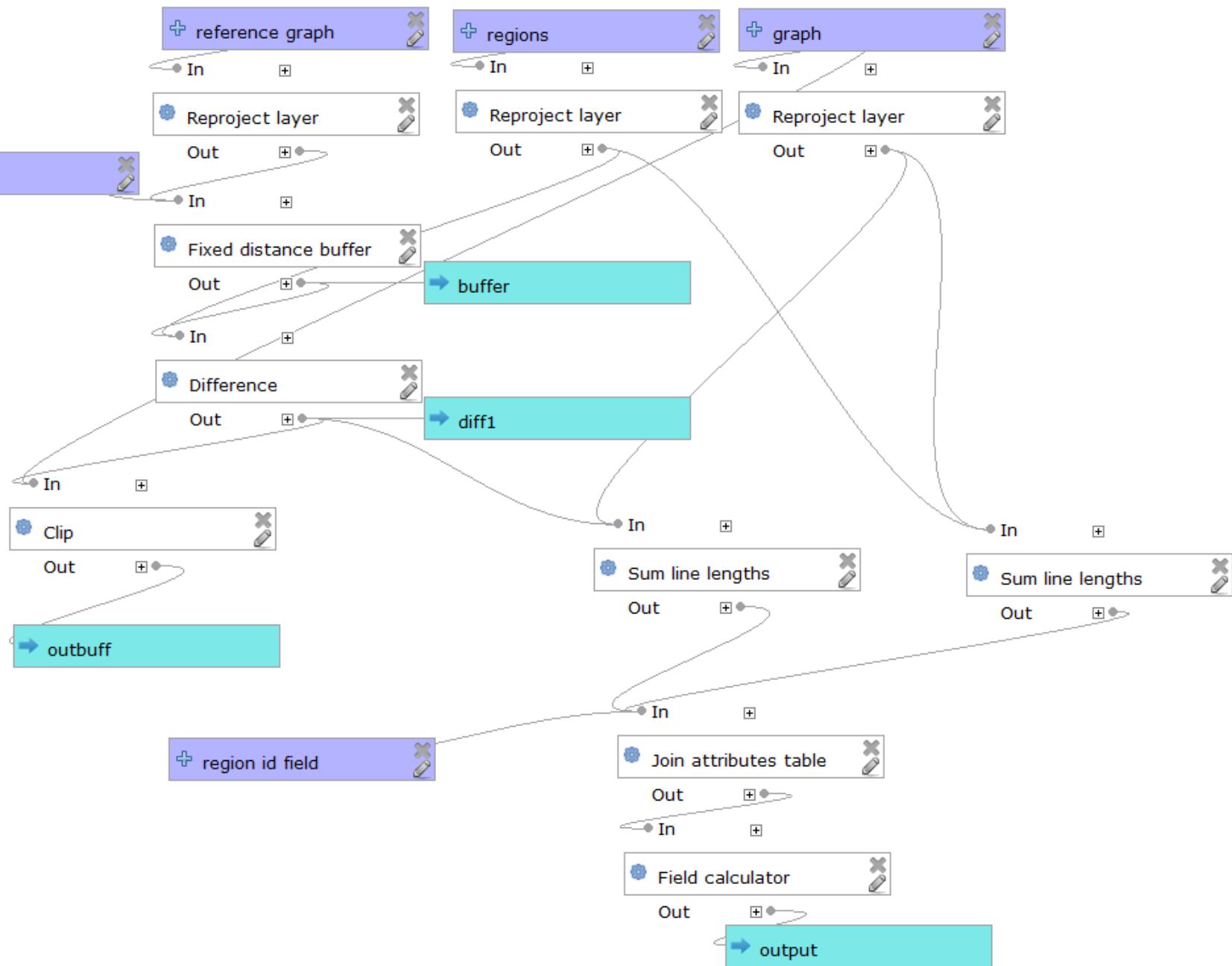










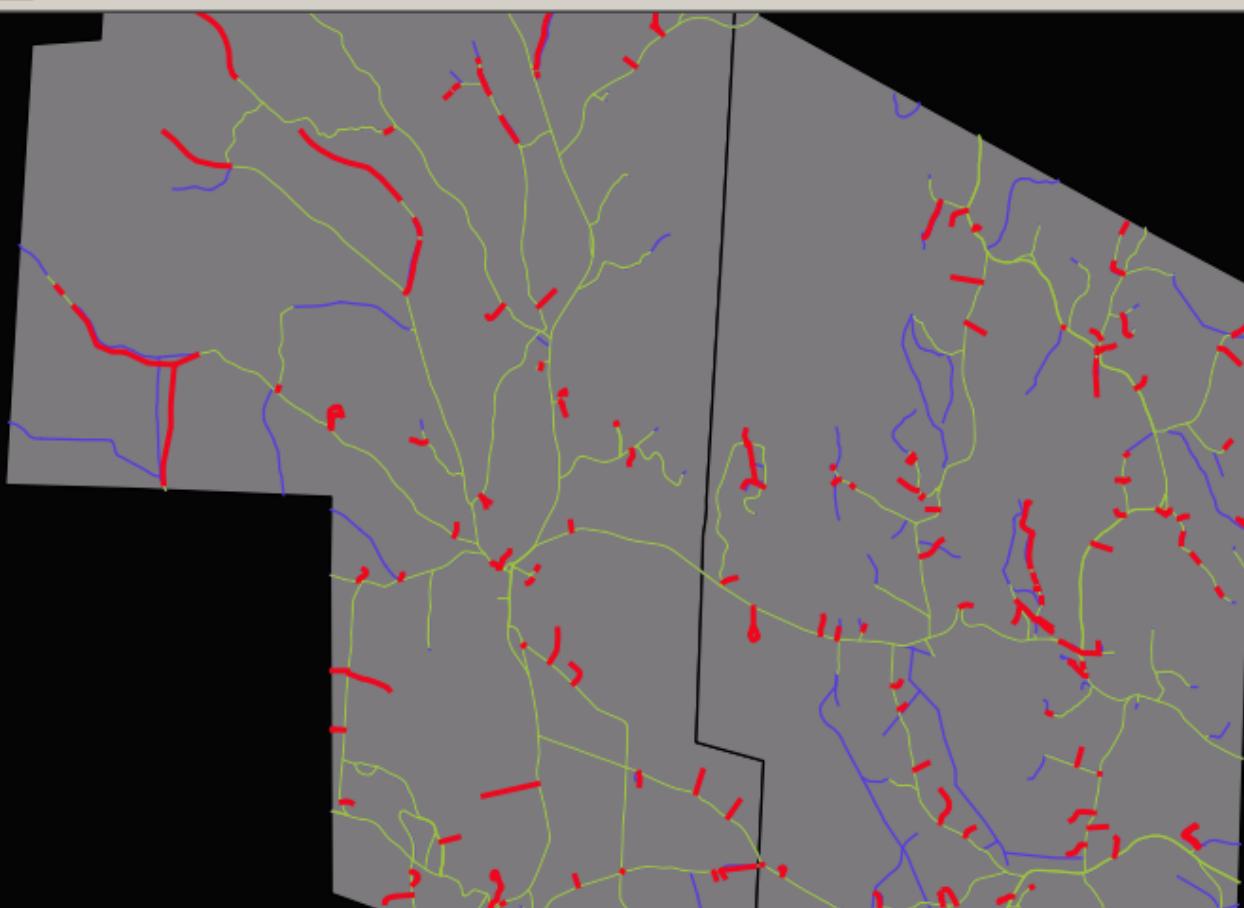
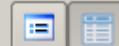


Attribute table - output :: Features total: 2, filtered: 2, selected: 0



	FIPS6	TOWNNAME	totalLEN	totalCNT	fips6_2	townname_2	outbufLEN	outbufCNT	outbuffPERC	
0	27105	WESTON	107555.05209...	125.00000000...	27105	WESTON	22215.438284...	68.000000000...	0.20655	
1	27005	ANDOVER	107008.61715...	161.00000000...	27005	ANDOVER	20440.991508...	102.000000000...	0.19102	

Show All Features ▾



Reference Dataset would have helped Apple here.



2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

E. Features are not duplicated.

CHECK: For a City, there should not be another City having the same name.

METHOD: Plugin: GroupStats

Q Group Stats

Data Features Window Help

	1	2
1	Function	count
2	NAME	
3	Valencia	2
4	Tripoli	2
5	San Jose	2
6	Hyderabad	2
7	Zibo	1
8	Zhengzhou	1
9	Zaragoza	1
10	Zahedan	1
11	Zagreb	1
12	Yuzhno-S...	1
13	Yokohama	1
14	Yerevan	1
15	Yellowknife	1
16	Yekaterinb...	1
17	Yazd	1
18	Yaounde	1
19	Yakutsk	1

Control panel

Layers
cities

Fields

- CAPITAL
- COUNTRY
- NAME
- POPULATION
- average
- count
- max
- median
- min

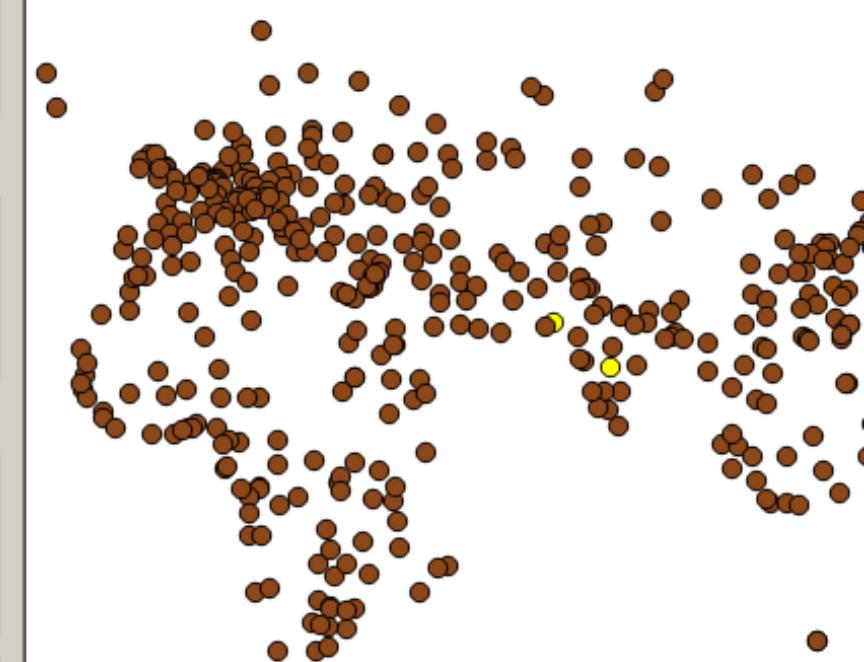
Filter Columns

	<input type="checkbox"/> count
--	--------------------------------

Rows Value use NULL values

<input type="checkbox"/> NAME	<input type="checkbox"/> NAME
-------------------------------	-------------------------------

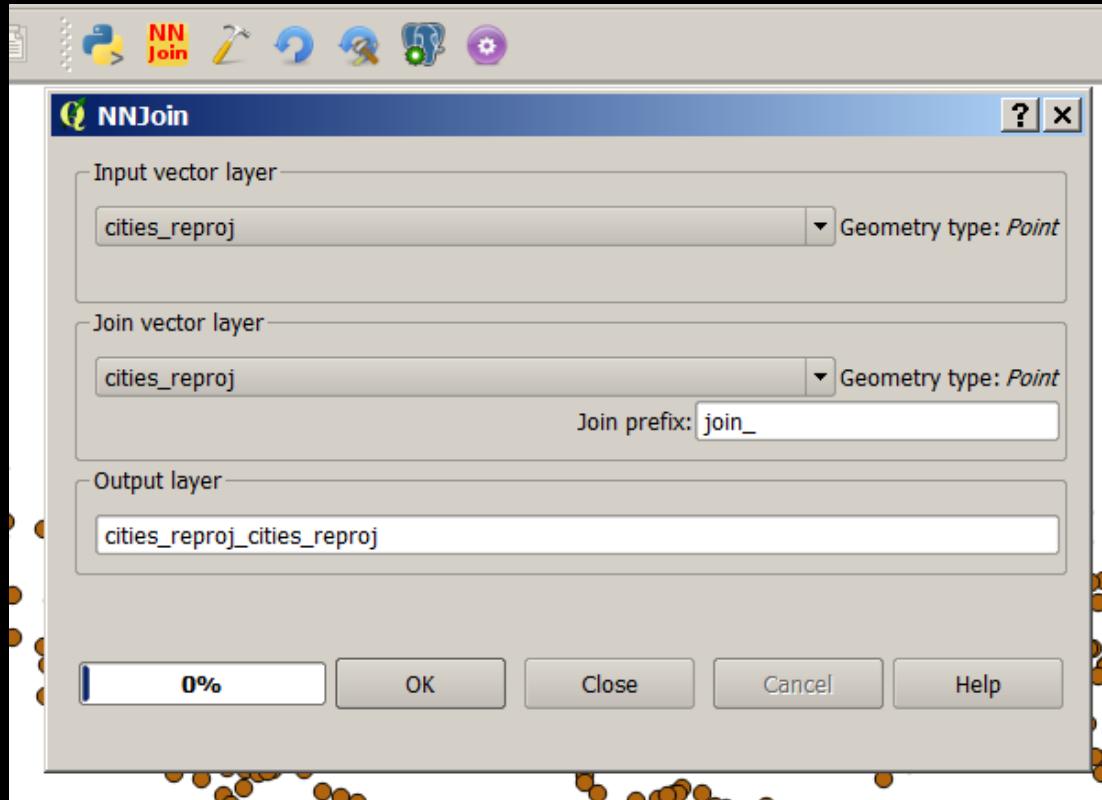
Use only selected features



E. Features are not duplicated.

CHECK: For a City, there should not be another City within X meters.

METHOD: Plugin: NNJoin



Input vector layer
cities_reproj Geometry type: Point

Join vector layer
cities_reproj Geometry type: Point
Join prefix: join_

Attribute table - cities_reproj_cities_reproj :: Features total: 606, filtered: 606, selected: 0

	NAME	COUNTRY	POPULATION	CAPITAL	join_NAME	join_COUNTRY	join_POPULATION	join_CAPITAL	distance
317	Callao	Peru	264133	N	Lima	Peru	4344000	Y	3091.47795110748
529	Lima	Peru	4344000	Y	Callao	Peru	264133	N	3091.47795110748
187	New Delhi	India	273036	Y	Delhi	India	7200000	N	5423.12745202839
188	Delhi	India	7200000	N	New Delhi	India	273036	Y	5423.12745202839
204	Taipei	Taiwan	6130000	Y	Chingmei	Taiwan	-99	N	7246.04162197582
354	Chingmei	Taiwan	-99	N	Taipei	Taiwan	6130000	Y	7246.04162197582
430	Fremantle	Australia	22484	N	Perth	Australia	994472	N	9052.9633540569
559	Perth	Australia	994472	N	Fremantle	Australia	22484	N	9052.9633540569
522	Khartoum	Sudan	924000	Y	Omdurman	Sudan	526287	N	9819.54544274401
552	Omdurman	Sudan	526287	N	Khartoum	Sudan	924000	Y	9819.54544274401
603	Hong Kong	UK	5395997	Y	Kowloon	UK	774781	N	11064.0800164803
604	Kowloon	UK	774781	N	Hong Kong	UK	5395997	Y	11064.0800164803
548	New York	US	16472000	N	Newark	US	329248	N	11972.4434858545
549	Newark	US	329248	N	New York	US	16472000	N	11972.4434858545

Input vector layer
cities_reproj

Join vector layer
cities_reproj

Attribute table - cities_reproj_cities_reproj

	NAME	COUNTRY	POPULA				PITAL	distance
317	Callao	Peru						3091.47795110748
529	Lima	Peru						3091.47795110748
187	New Delhi	India						5423.12745202839
188	Delhi	India	7200000	N				5423.12745202839
204	Taipei	Taiwan	6130000	Y	Chingmei	Taiwan		7246.04162197582
354	Chingmei	Taiwan	-99	N	Taipei	Taiwan	6130000	Y
430	Fremantle	Australia	22484	N	Perth	Australia	994472	N
559	Perth	Australia	994472	N	Fremantle	Australia	22484	N
522	Khartoum	Sudan	924000	Y	Omdurman	Sudan	526287	N
552	Omdurman	Sudan	526287	N	Khartoum	Sudan	924000	Y
603	Hong Kong	UK	5395997	Y	Kowloon	UK	774781	N
604	Kowloon	UK	774781	N	Hong Kong	UK	5395997	Y
548	New York	US	16472000	N	Newark	US	329248	N
549	Newark	US	329248	N	New York	US	16472000	N
								11972.4434858545

Layer needs to be projected (not on the fly) for useful distance measurements

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.



2. Specs and Checks

- A. City population values are realistic.
- B. Road attribute values are on a defined list.
- C. Roads are captured within 10m of their locations in reality.
- D. Captured roads exist in reality.
- E. Features are not duplicated.

A Survey of QGIS Methods for Detecting Data Quality Errors

1

- The opening question:
- *What are you looking for?*

2

- Examples:
- *Specs and Checks*

3

- The closing question:
- *What is your framework?*

3. What is your framework?

- Ad hoc or suite of checks?
 - model of models
- How many layers will you checks?
 - batch processing ([link](#))
- How will you solve the violations?
 - be cautious with automated solving = “could make things irretrievably worse”
- Are there false positives?
- How will you store violations?
 - problems found, actions taken, false positive persistence
- Incorporate into data versioning system (ie. geogig)?
- What are the useful metrics or reports?

3. What is your framework?

Resources

Setting up a framework:

- <http://stats.stackexchange.com/questions/11659/essential-data-checking-tests/11669#11669>
- <http://stats.stackexchange.com/questions/7467/quality-assurance-and-quality-control-qa-qc-guidelines-for-a-database/7472#7472>

Examples of application from ESRI:

- <http://www.esri.com/library/whitepapers/pdfs/gis-data-quality-best-practices.pdf>

Theory meets practice, from a GIS instructor:

- <http://www.nuim.ie/staff/dpringle/gis/gis11.pdf>

3. What is your framework?

Resources

Good podcast about QGIS, interviews with developers:

- <http://qgispodcast.libsyn.com/>

Q&A in the "data quality" tag at GIS.SE

- <http://gis.stackexchange.com/questions/tagged/data-quality>

Incorporating R:

- <http://arc-team-open-research.blogspot.com.br/2013/01/manager-usefull-plugin-for-qgis.html>

Spatial statistics:

- https://docs.qgis.org/2.8/en/docs/training_manual/vector_analysis/spatial_statistics.html

A Survey of QGIS Methods for Detecting Data Quality Errors

1

- The opening question:
- *What are you looking for?*

2

- Examples:
- *Specs and Checks*

3

- The closing question:
- *What is your framework?*

