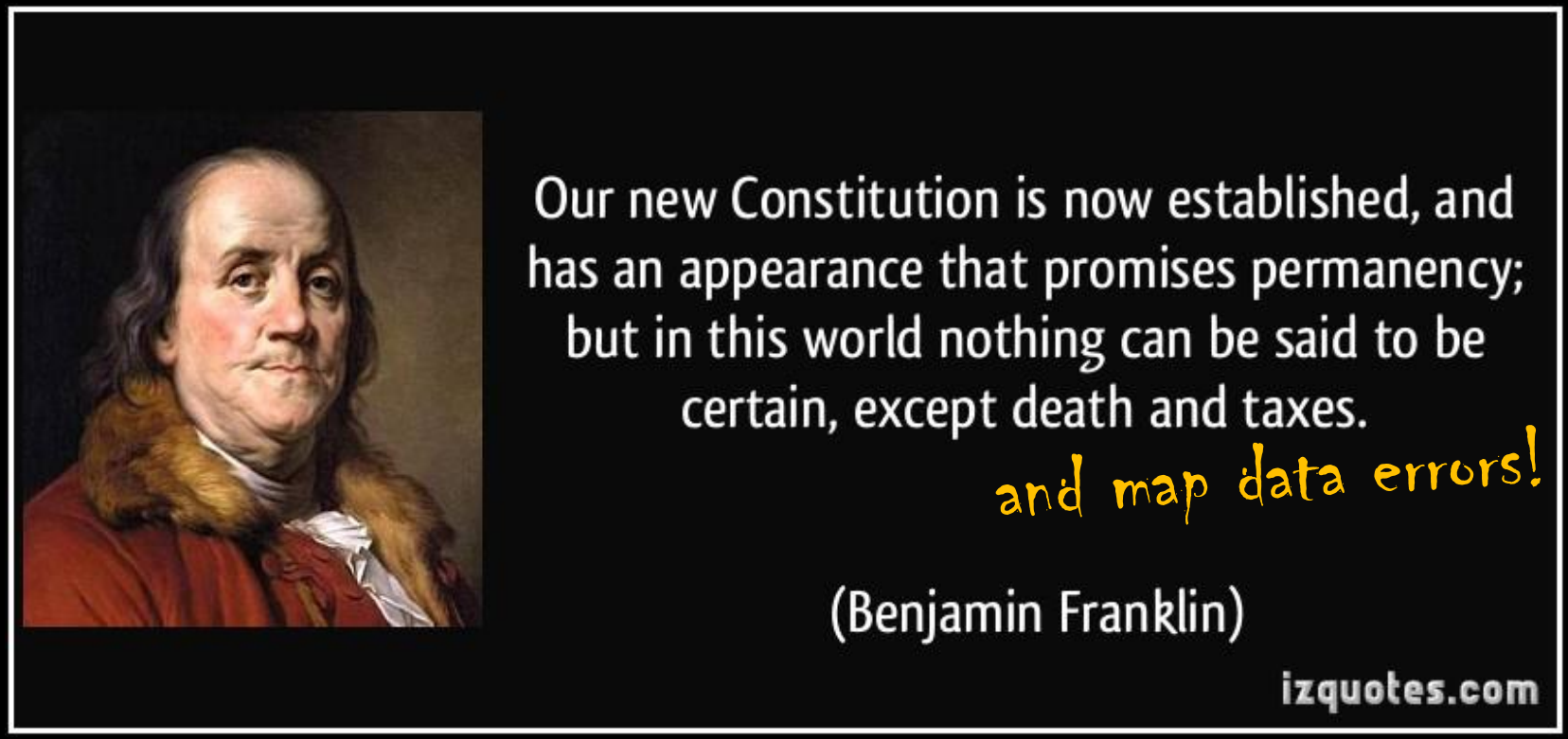
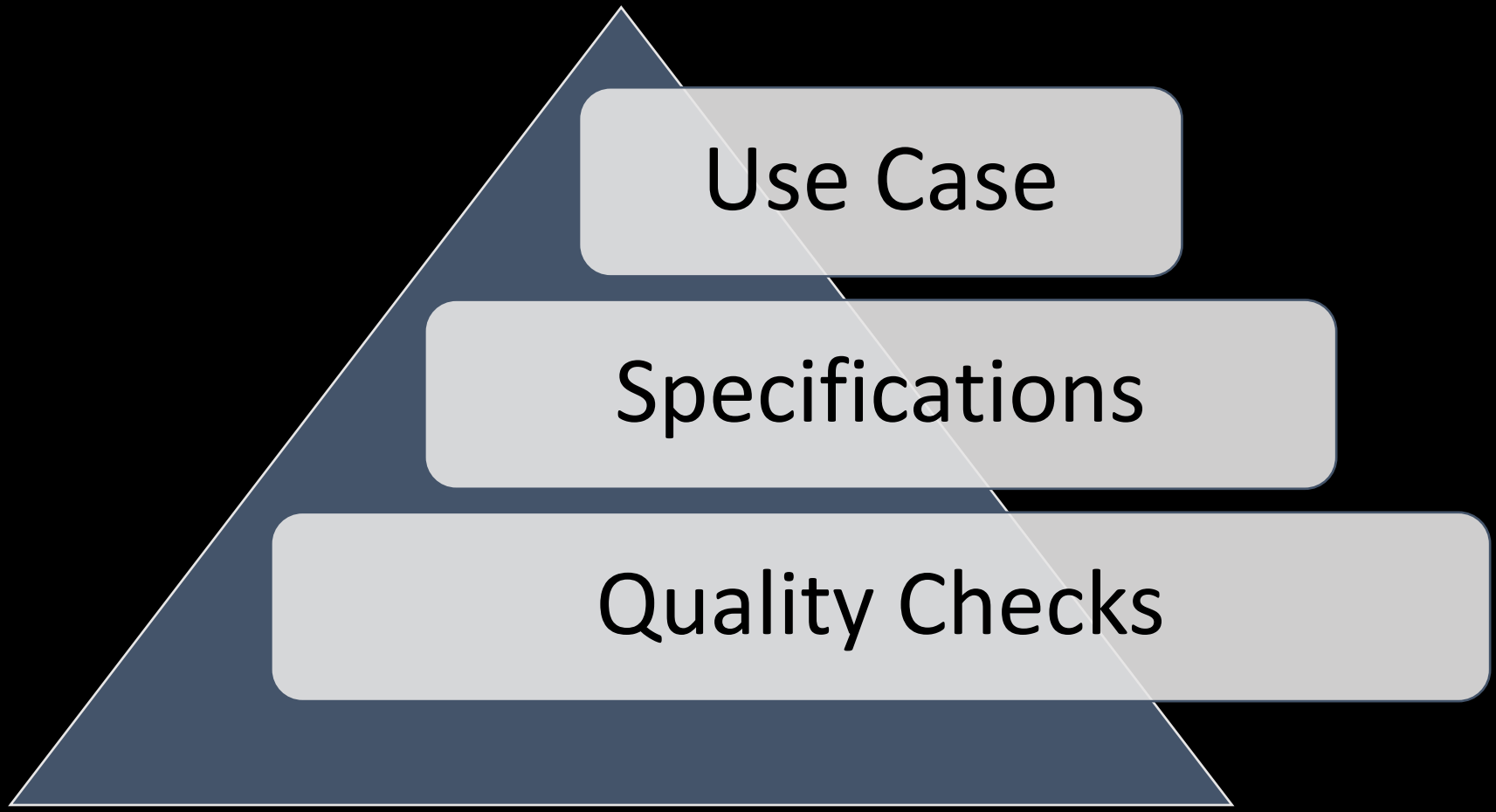
A Survey of QGIS Methods for Detecting Data Quality Errors



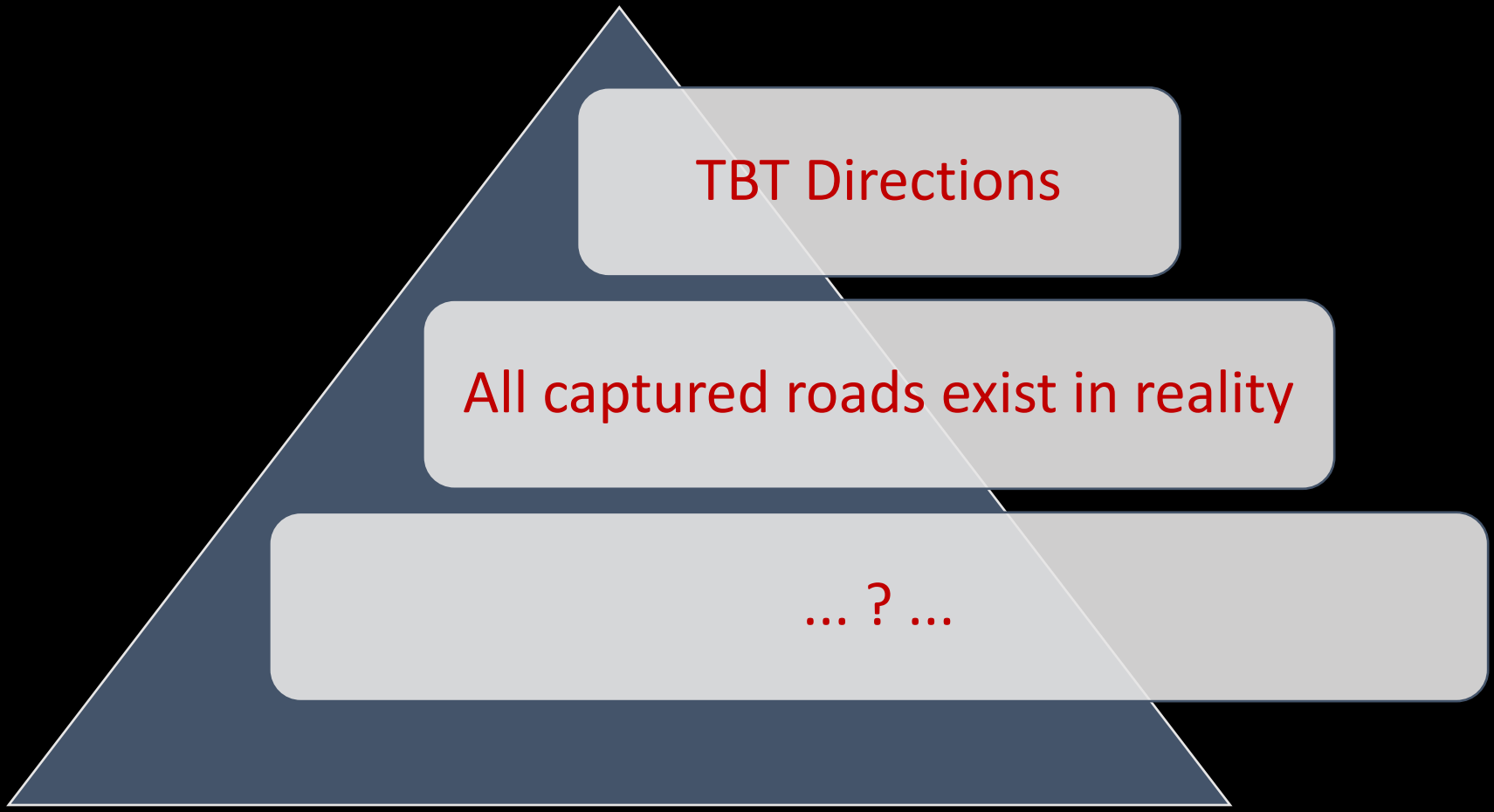
A Survey of QGIS Methods for Detecting Data Quality Errors

1	• The opening question: • ***What are you looking for?***
2	• Examples: • ***Specs and Checks***
3	• The closing question: • ***What is your framework?***

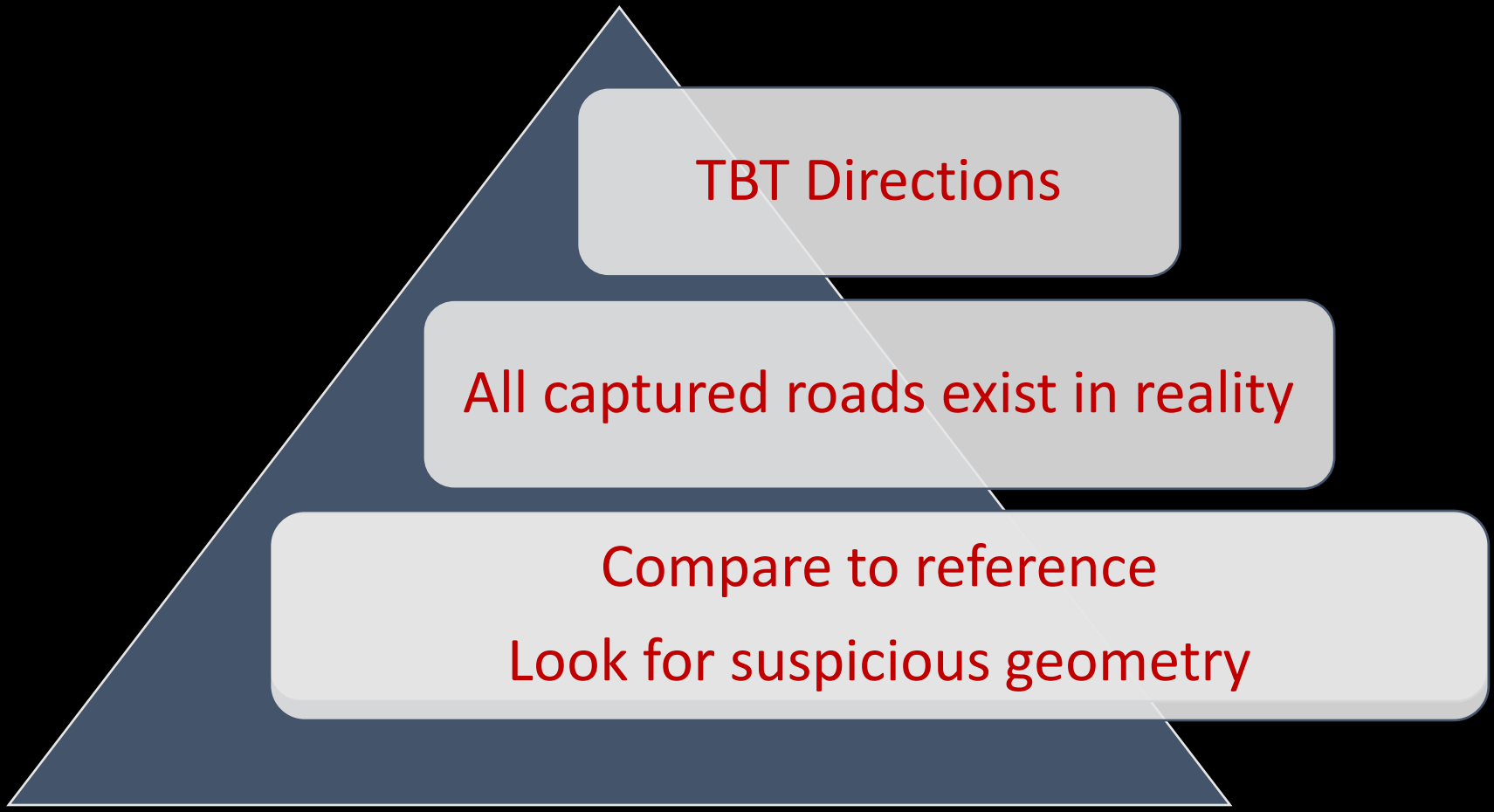
1. What are you looking for?



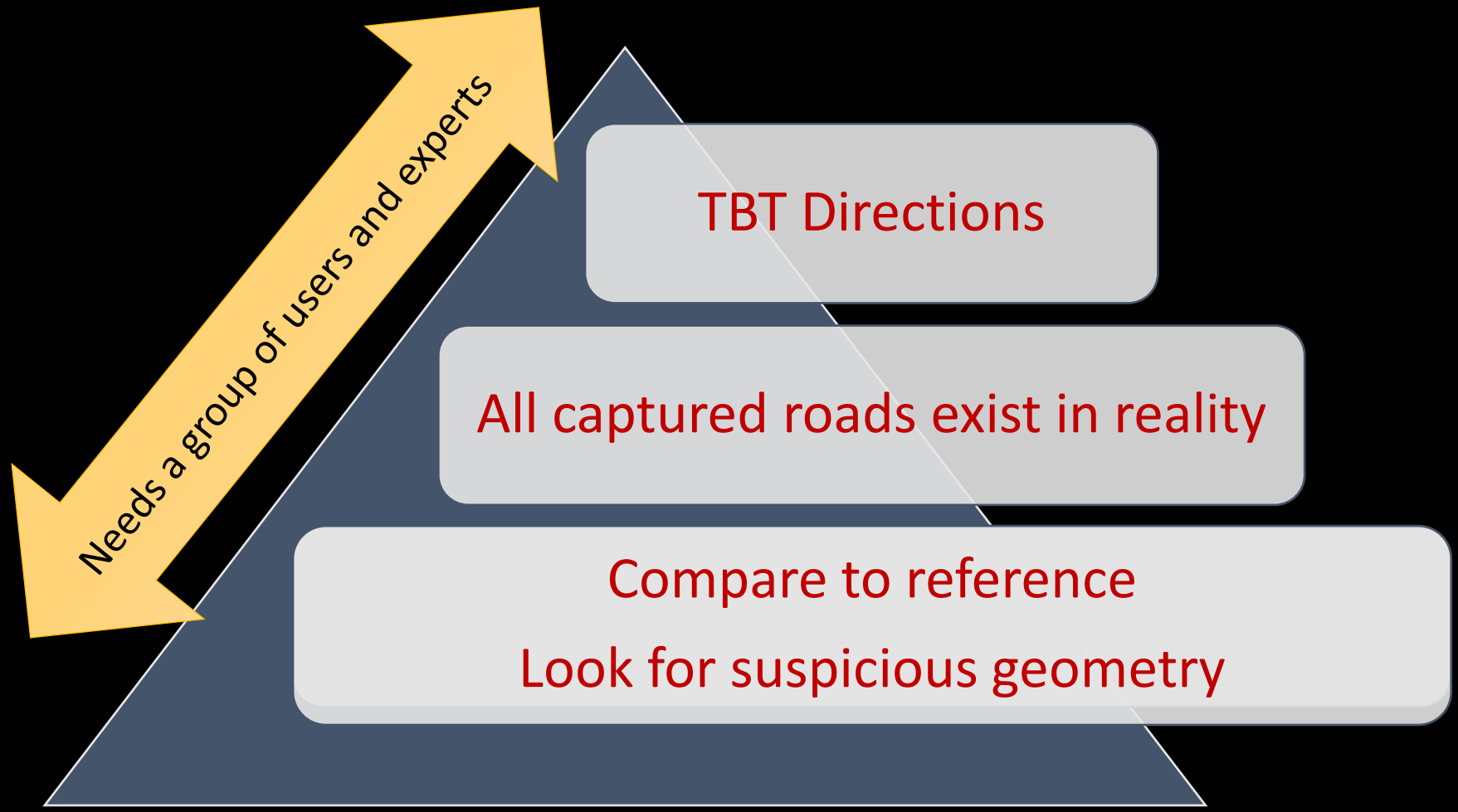
1. What are you looking for?



1. What are you looking for?



1. What are you looking for?



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): I
with Google, Apple decided to get rid of the much-adored Go
app that had always come on new iPhones—and to replace it
map app that Apple had written itself

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

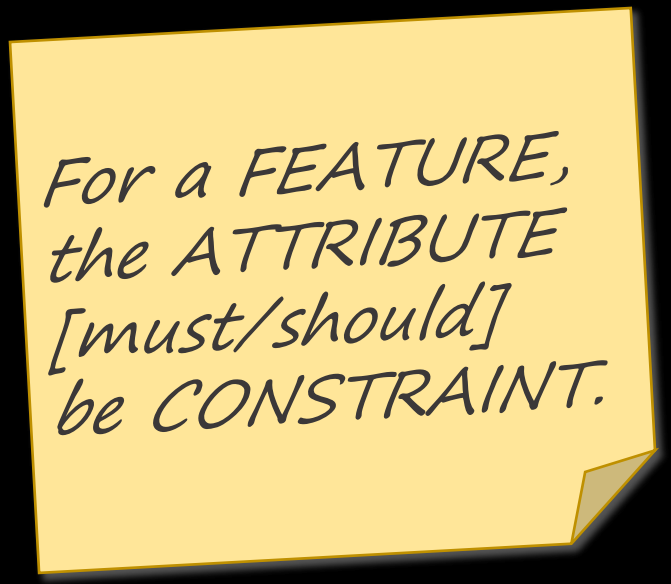
A. City population values are realistic.
B. Road attribute values are on a defined list.
C. Roads are captured within 10m of their locations in reality.
D. Captured roads exist in reality.
E. Features are not duplicated.

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.

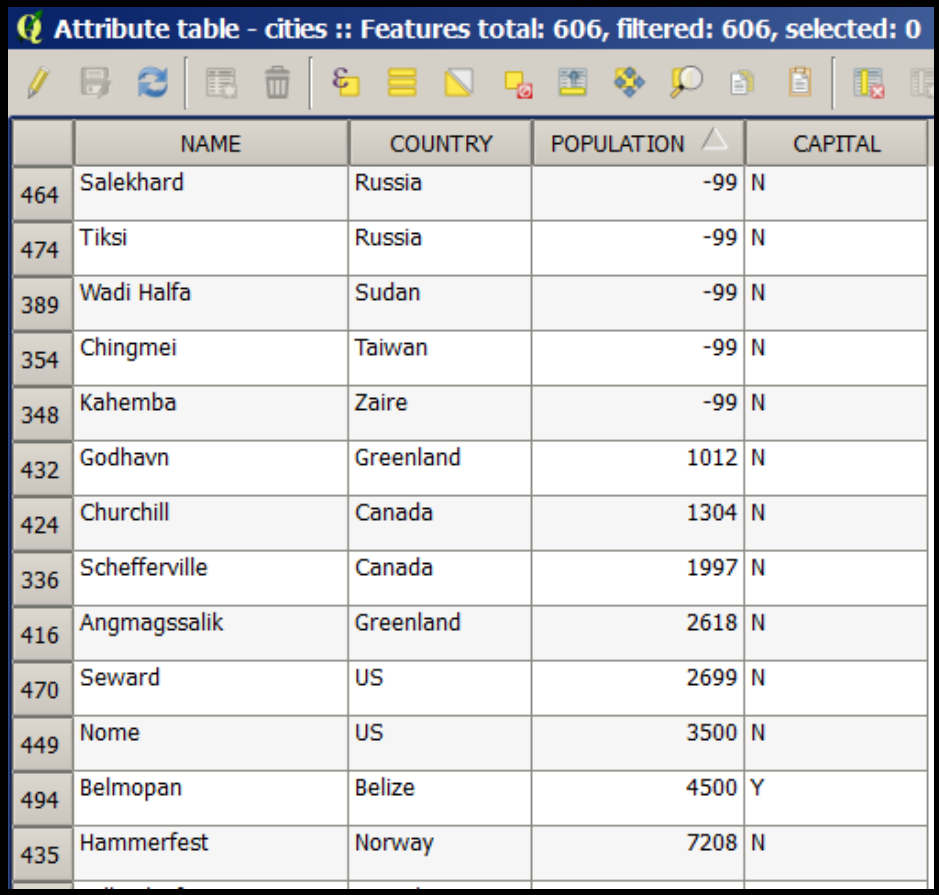


A. City population values are realistic.

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.



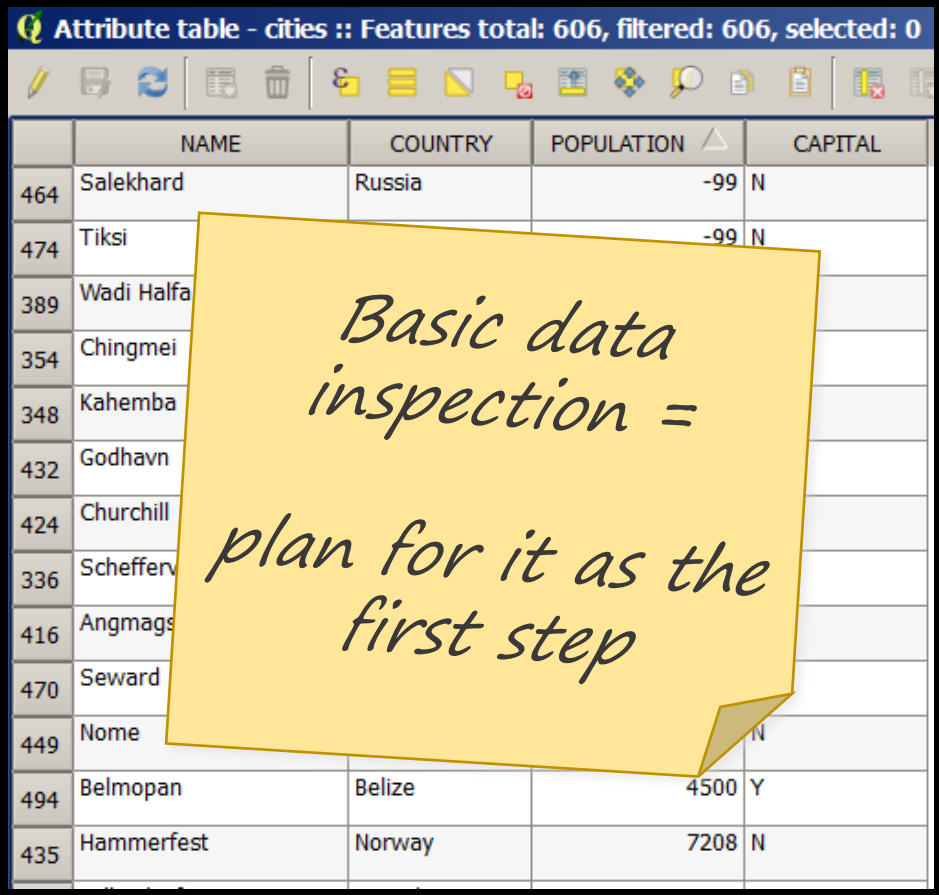
	NAME	COUNTRY	POPULATION	CAPITAL
464	Salekhard	Russia	-99	N
474	Tiksi	Russia	-99	N
389	Wadi Halfa	Sudan	-99	N
354	Chingmei	Taiwan	-99	N
348	Kahemba	Zaire	-99	N
432	Godhavn	Greenland	1012	N
424	Churchill	Canada	1304	N
336	Schefferville	Canada	1997	N
416	Angmagssalik	Greenland	2618	N
470	Seward	US	2699	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.



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:

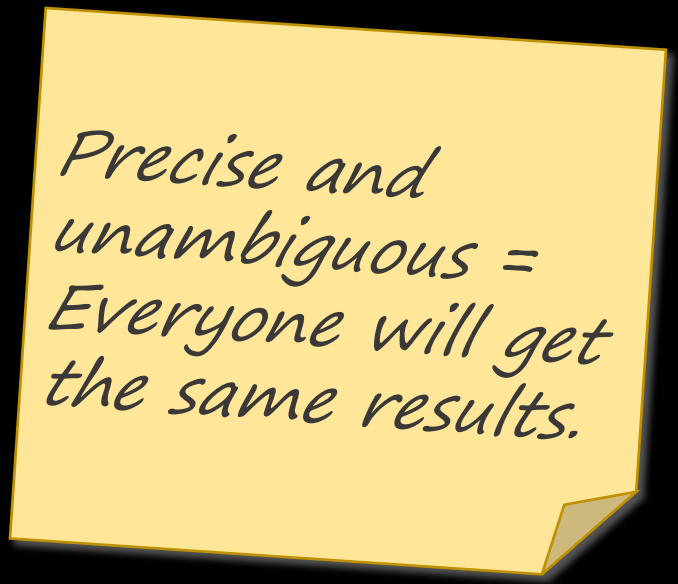
Select by Expression

METHOD 2:

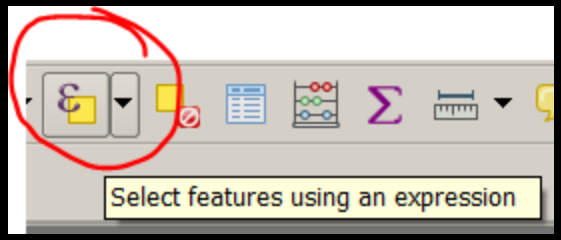
Custom Expression

METHOD 3:

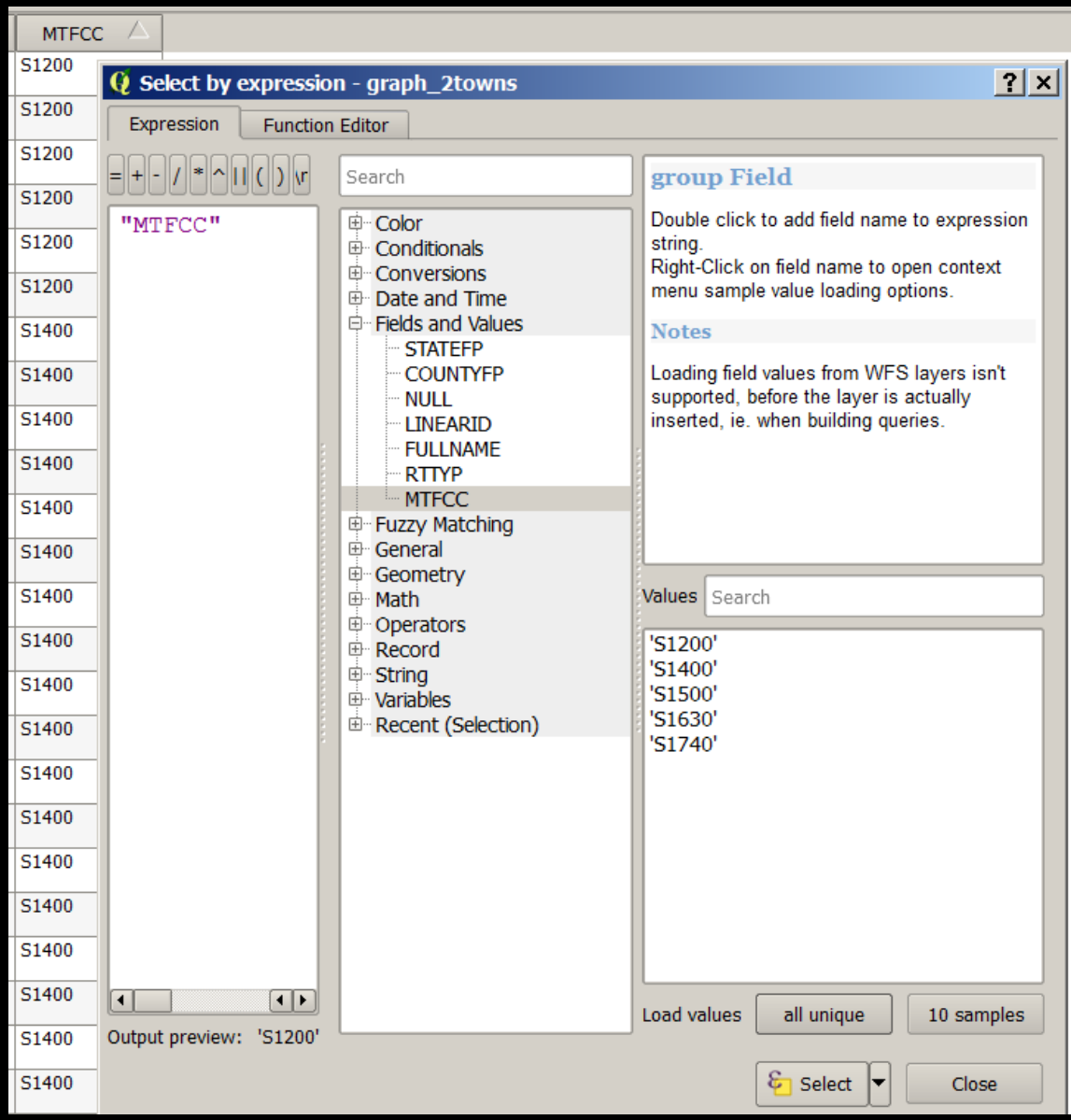
Plugin: GroupStats



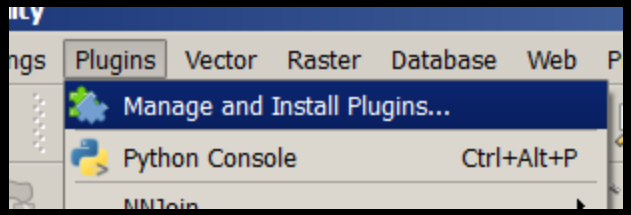
METHOD 1

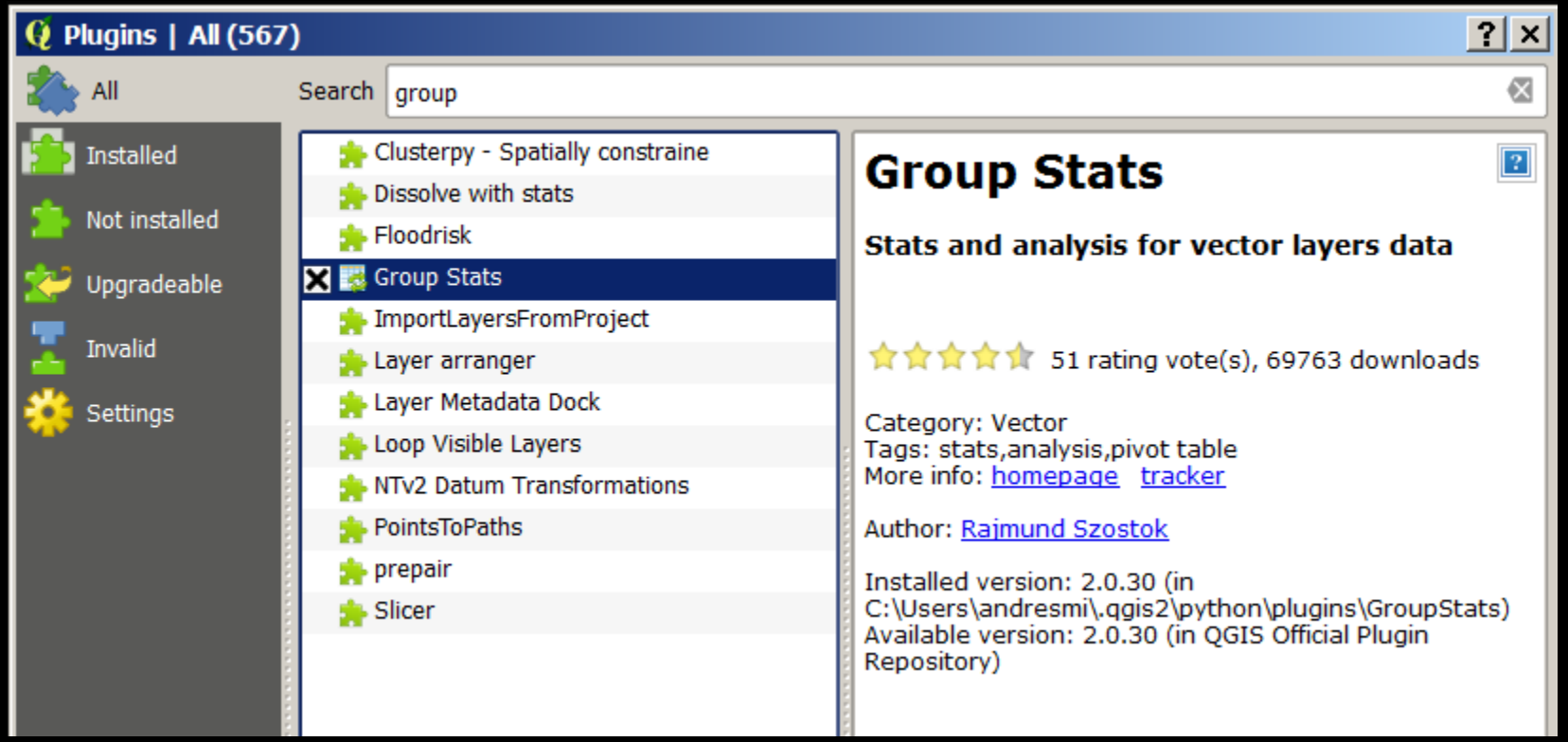


- Select By Expression
- Fields and Values
- Load values: all unique
- ... finish expression with any invalid values

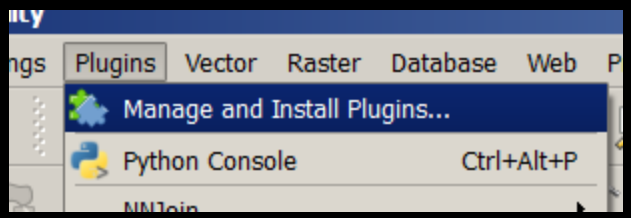


METHOD 2

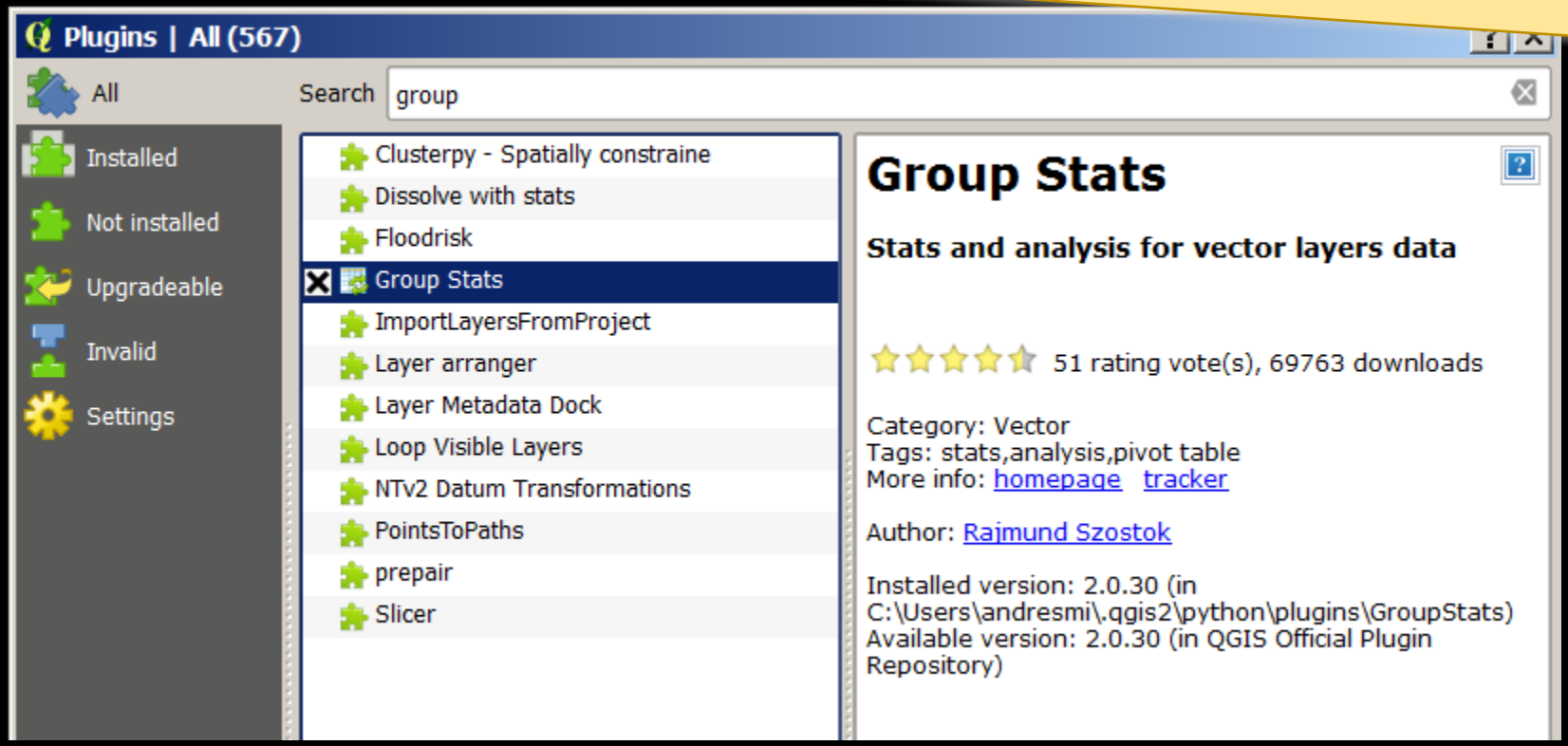

Plugins
Vector
Raster
Database
Web
Manage and Install Plugins...
Python Console
Ctrl+Alt+P


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)

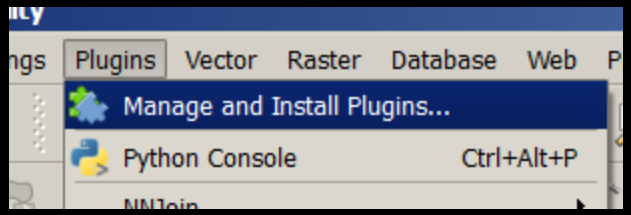
METHOD 2



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

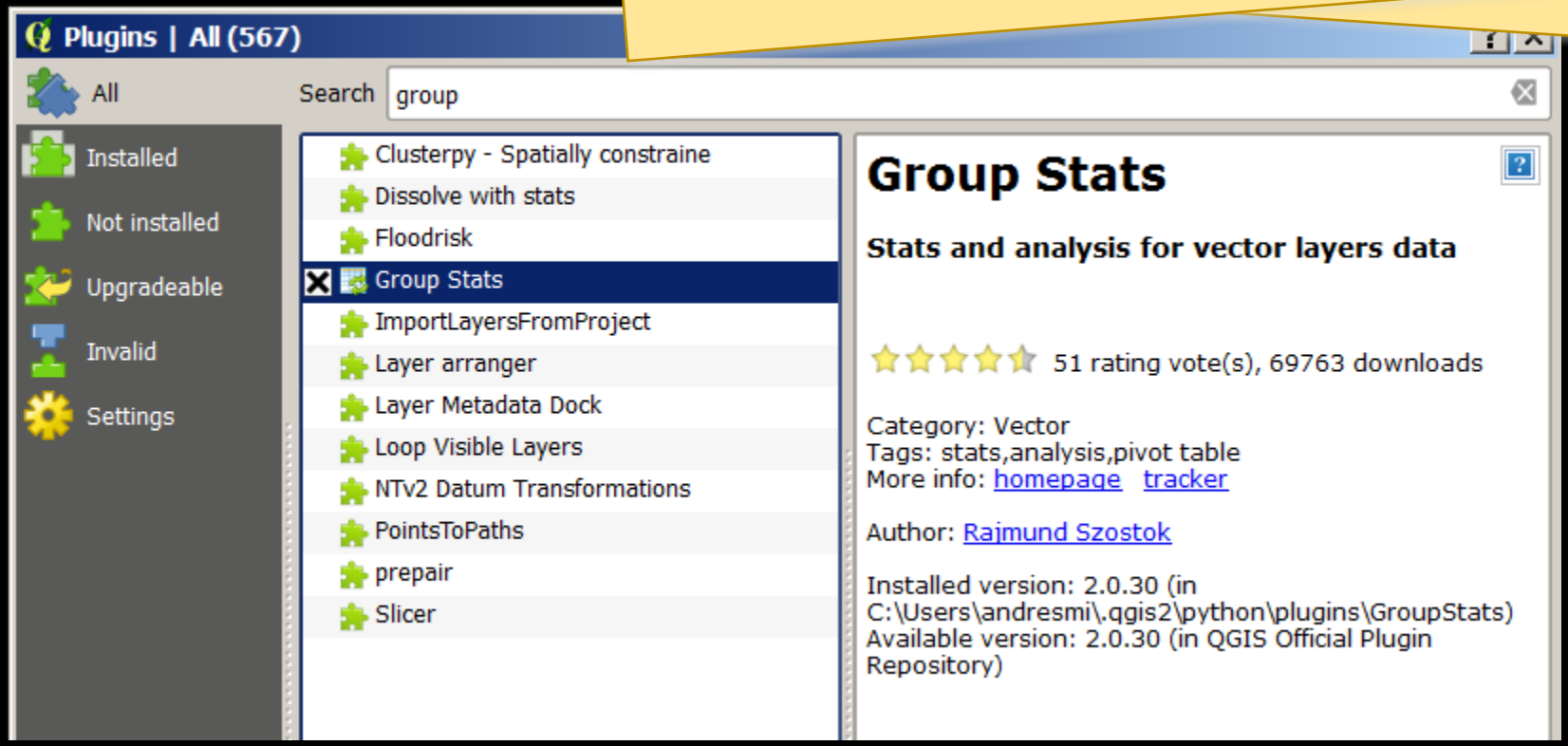


METHOD 2

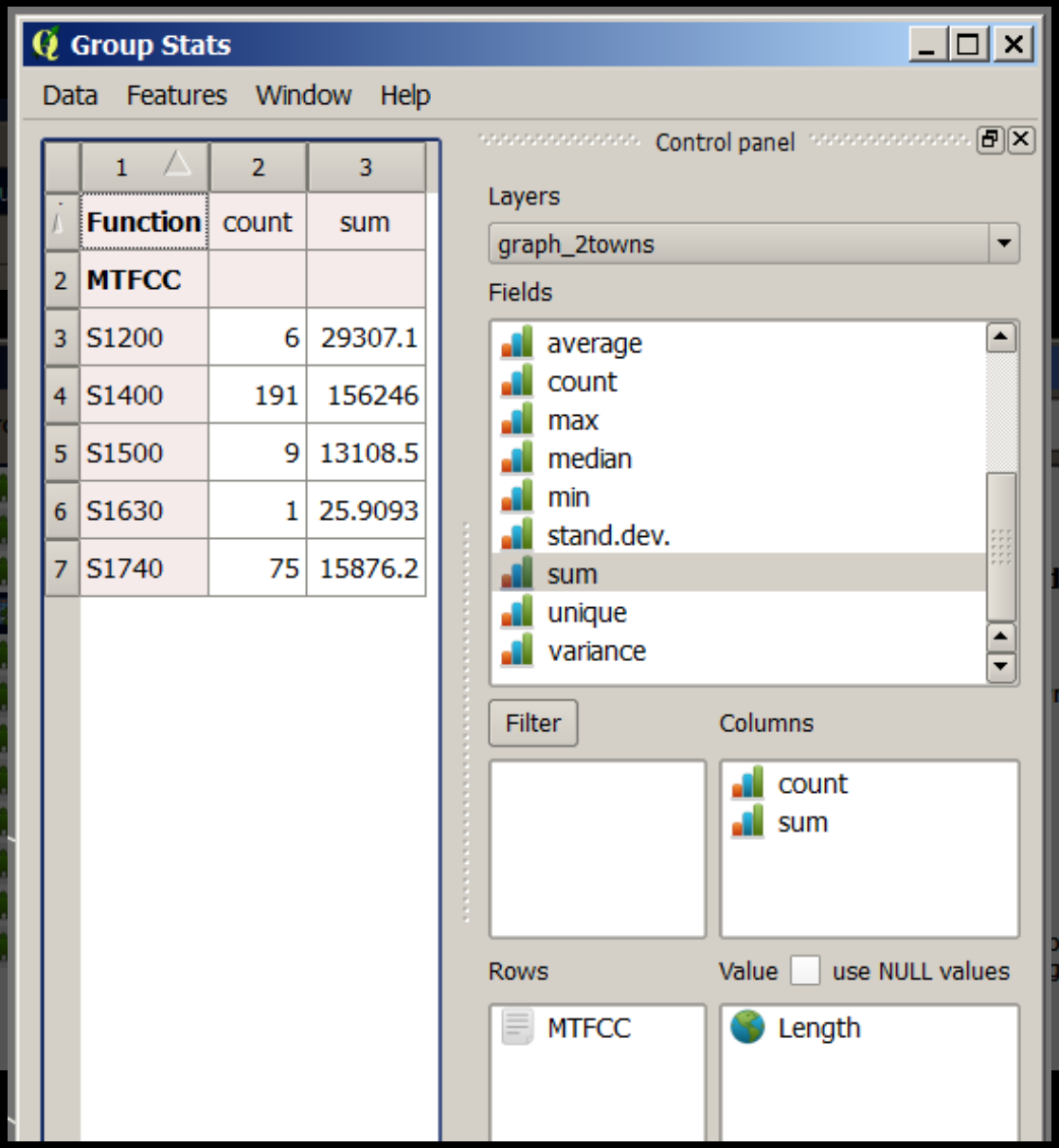


Or build your own plugin:

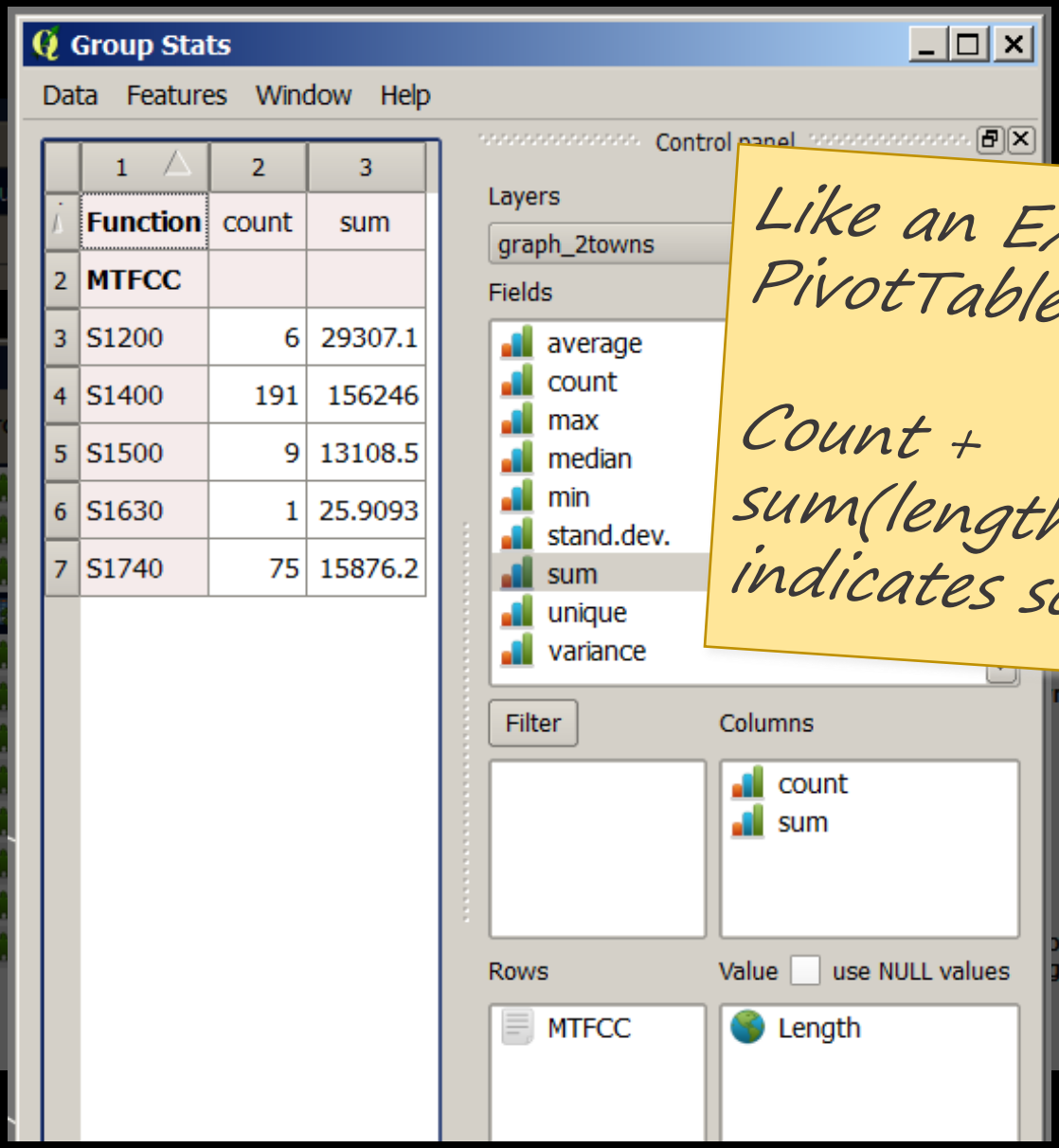
http://www.qgistutorials.com/en/docs/building_a_python_plugin.html



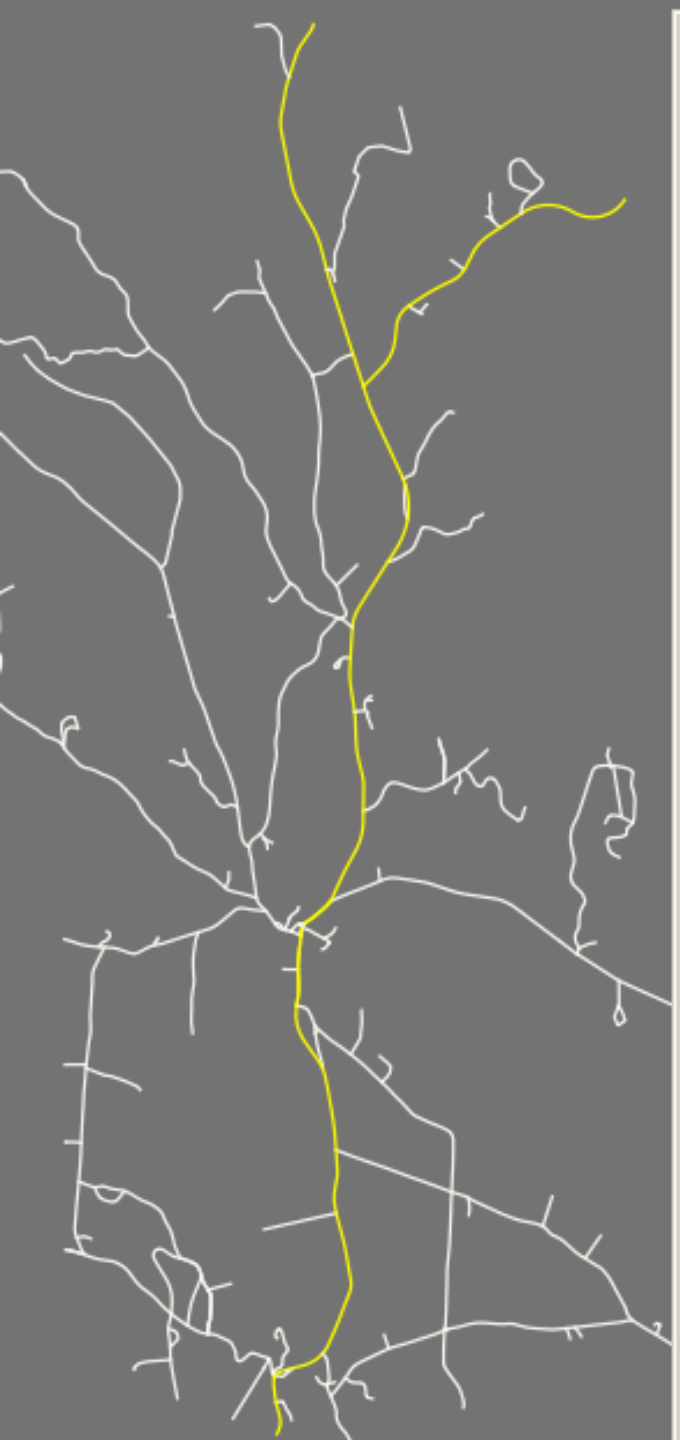
METHOD 2


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

METHOD 2



	1	2	3
1	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



Group Stats

Data Features Window Help

	1	2	3
1	**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

Control panel

Layers

← click a row then Features > show selected on map

variance

Filter

Columns

count
sum

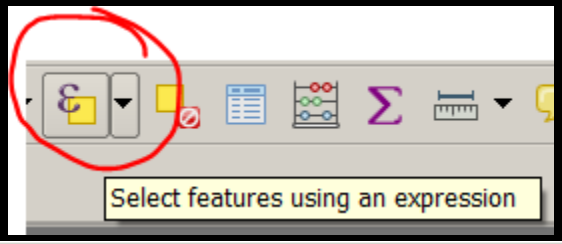
Rows

MTFCC

Value use NULL values

Length

METHOD 3



Select by expression - graph_2towns

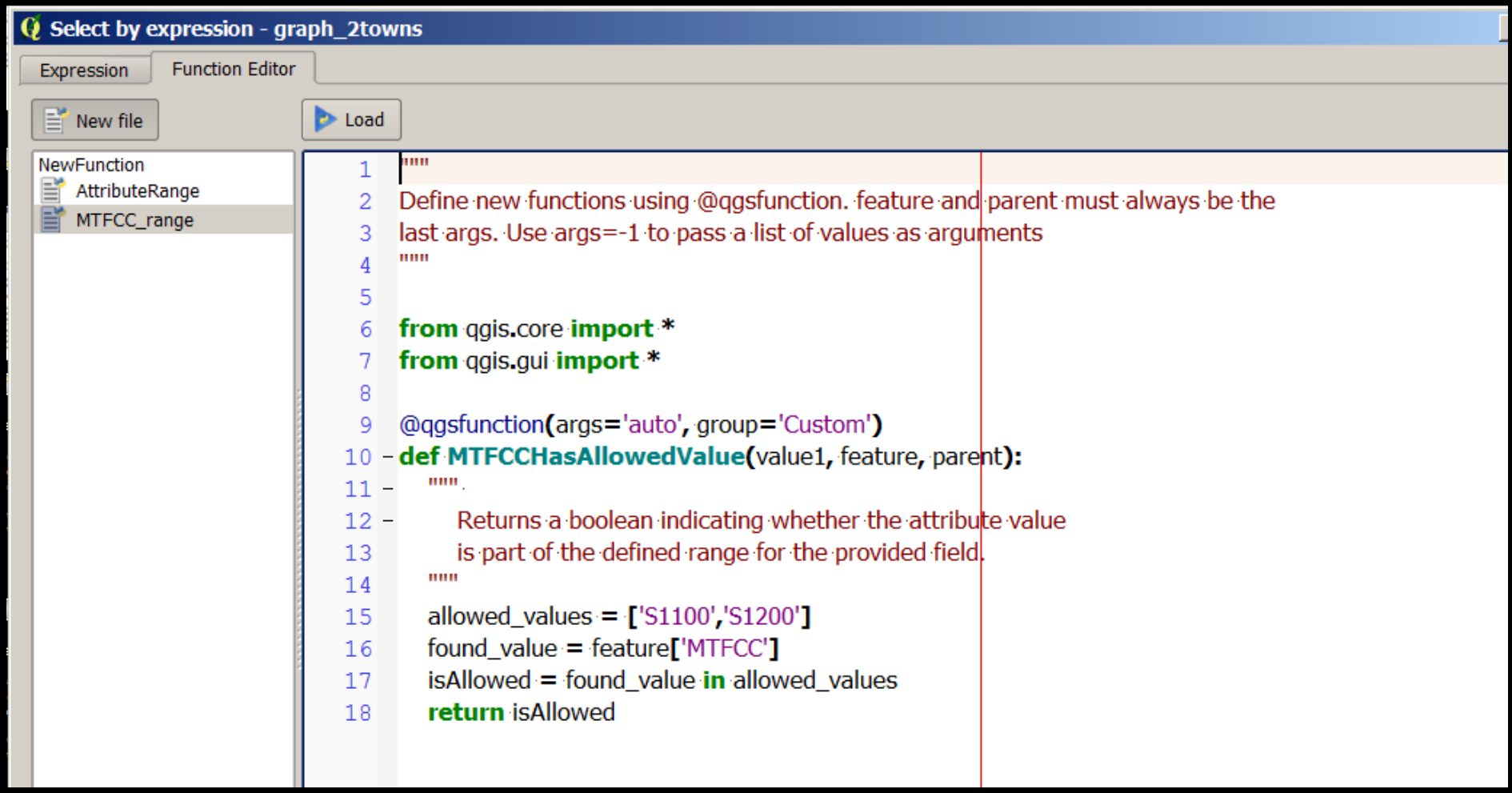
Expression | Function Editor

New file | Load

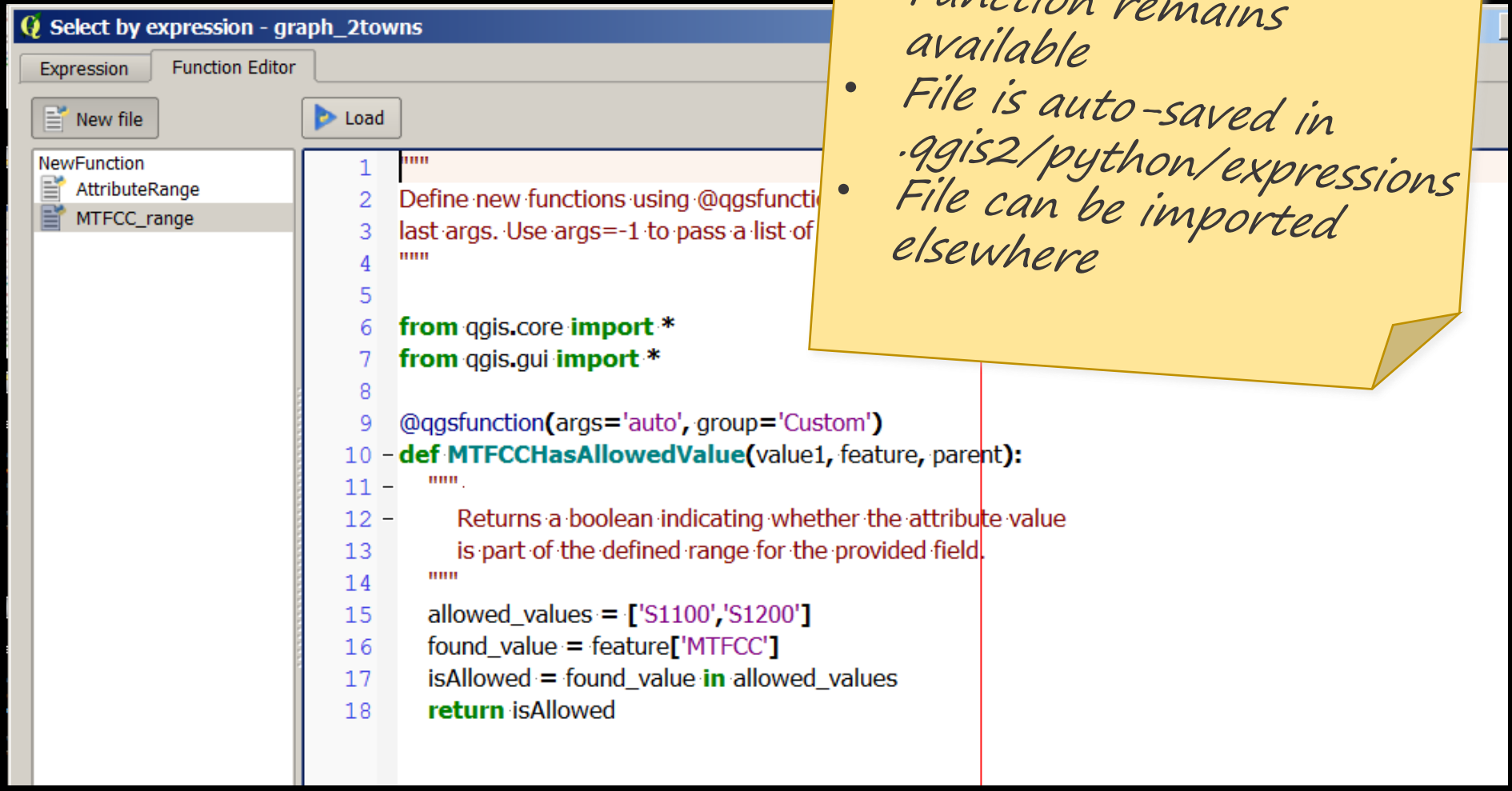
NewFunction
- AttributeRange
- MTFCC_range

```
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 func(value1, feature, parent):
11     return value1
12
```

METHOD 3

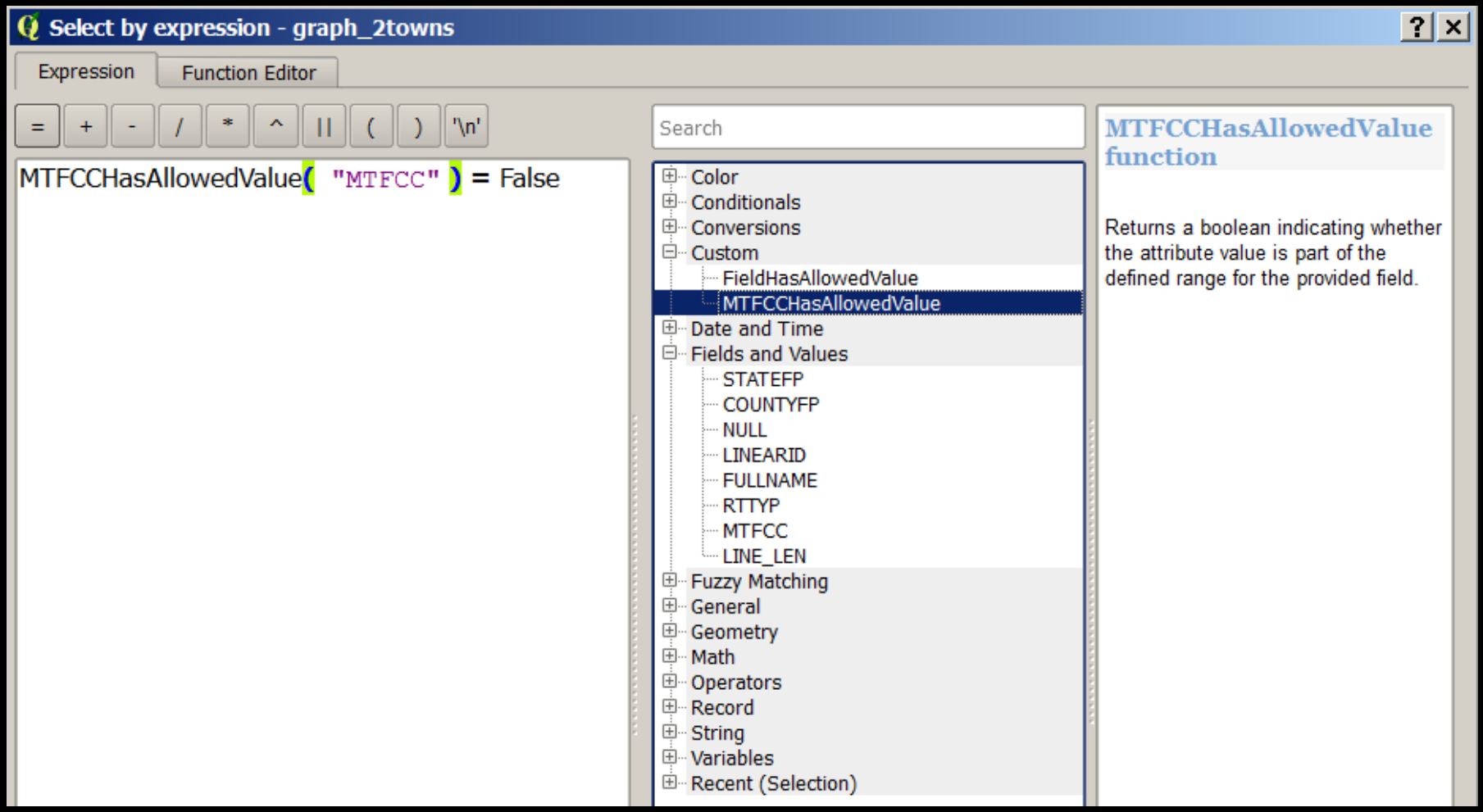

Select by expression - graph_2towns
Expression
Function Editor
New file
Load
NewFunction
AttributeRange
MTFCC_range
"""
Define new functions using @qgsfunction. feature and parent must always be the
last args. Use args=-1 to pass a list of values as arguments
"""
from qgis.core import *
from qgis.gui import *
@qgsfunction(args='auto', group='Custom')
def MTFCCHasAllowedValue(value1, feature, parent):
"""
Returns a boolean indicating whether the attribute value
is part of the defined range for the provided field.
"""
allowed_values = ['S1100','S1200']
found_value = feature['MTFCC']
isAllowed = found_value in allowed_values
return isAllowed

METHOD 3

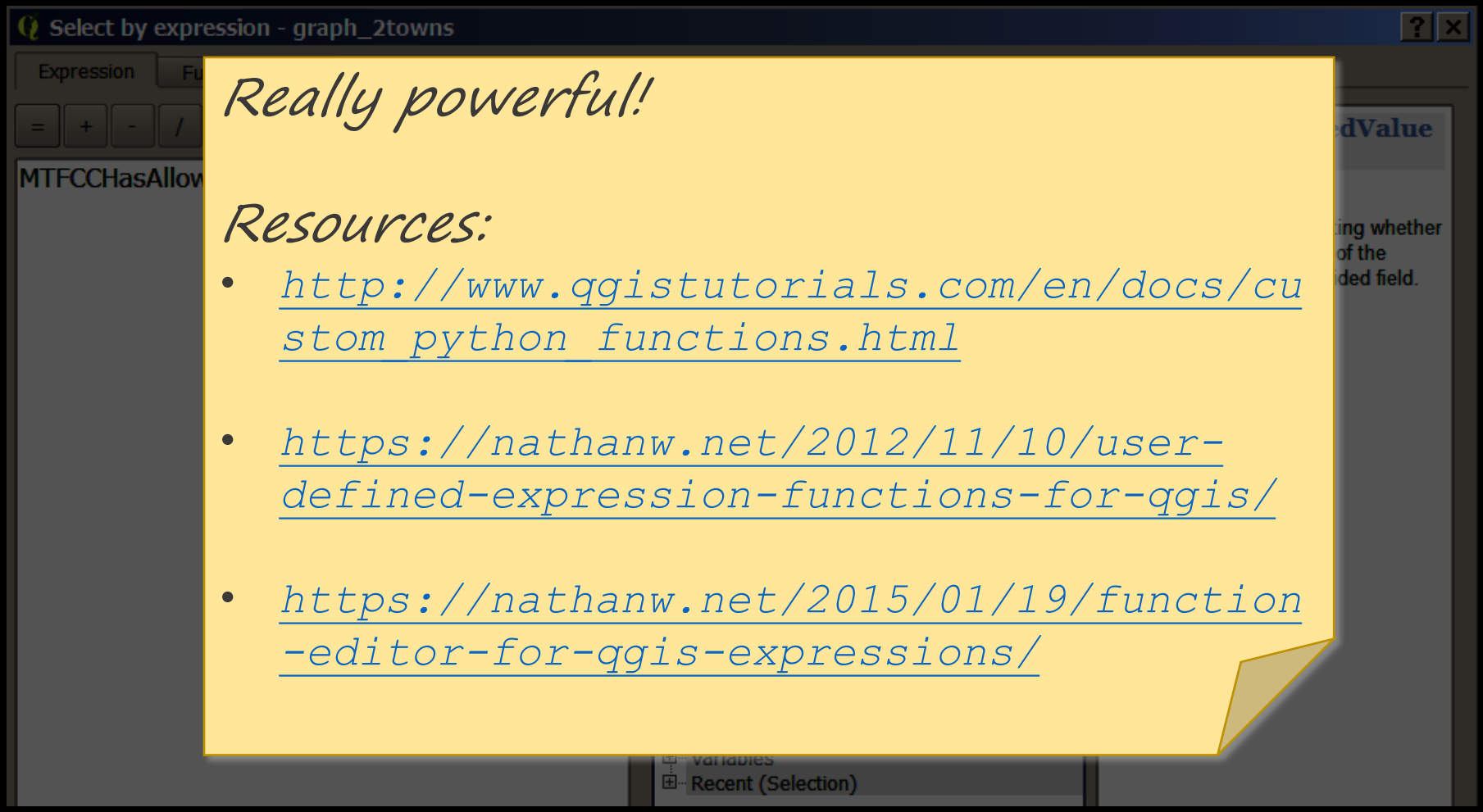


- Click Load
- Function remains available
- File is auto-saved in .qgis2/python/expressions
- File can be imported elsewhere

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*
- *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

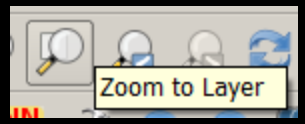
A. City population values are realistic.
B. Road attribute values are on a defined list.
C. Roads are captured within 10m of their locations in reality.
D. Captured roads exist in reality.
E. Features are not duplicated.

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

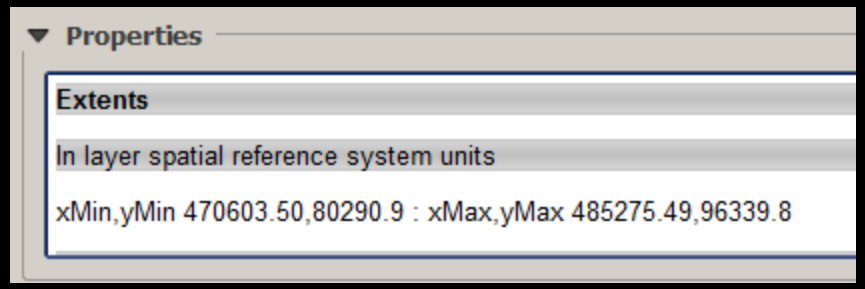


METHOD 2:

Layer Properties >

Metadata >

Extents

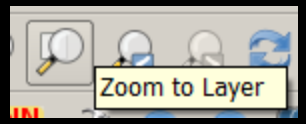


C. Roads are captured within 10m of their locations in reality.

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

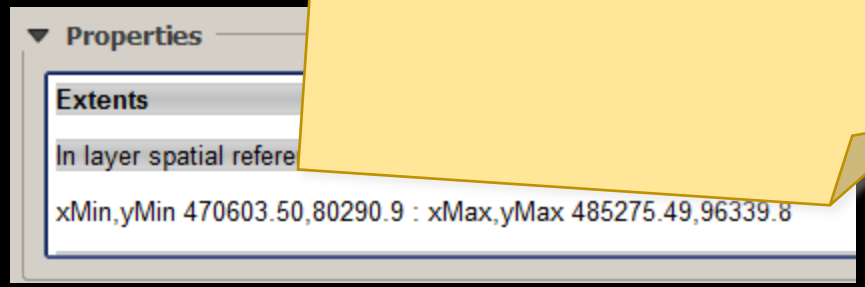


METHOD 2:

Layer Properties >

Metadata >

Extents

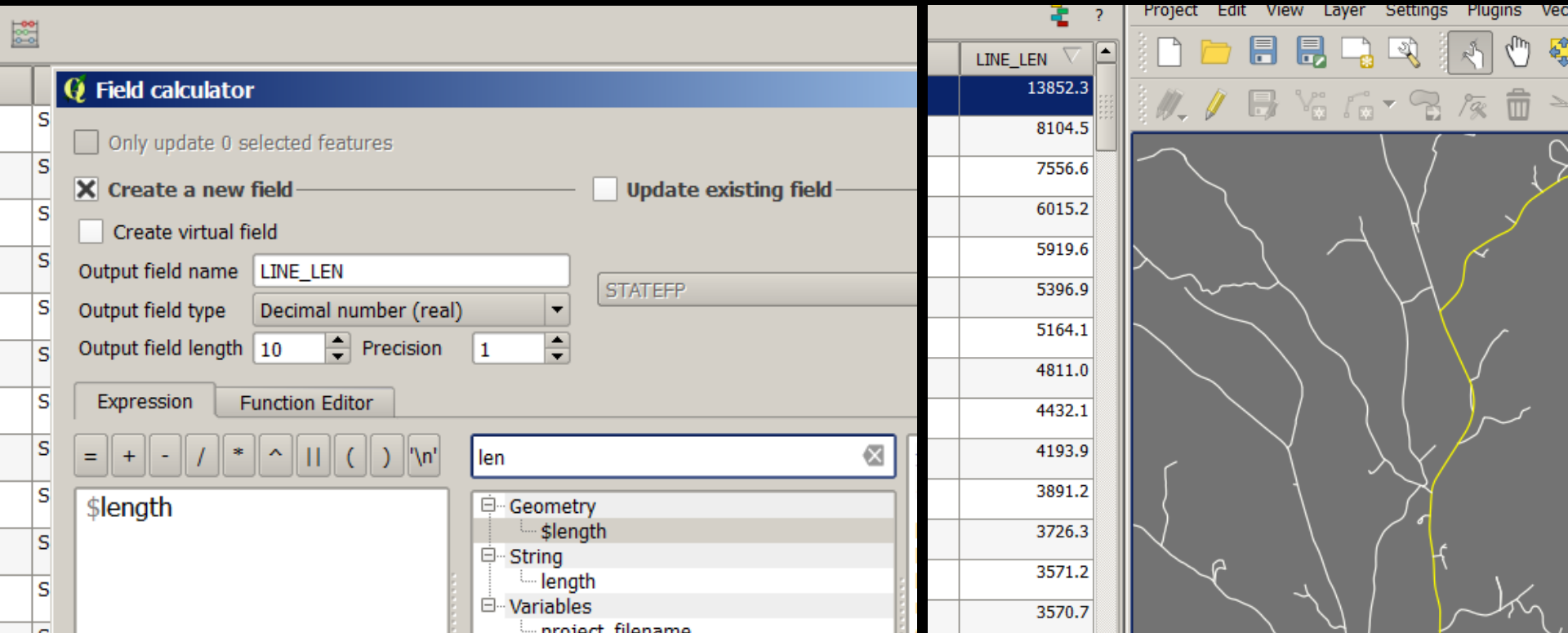


More basic data inspection

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

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

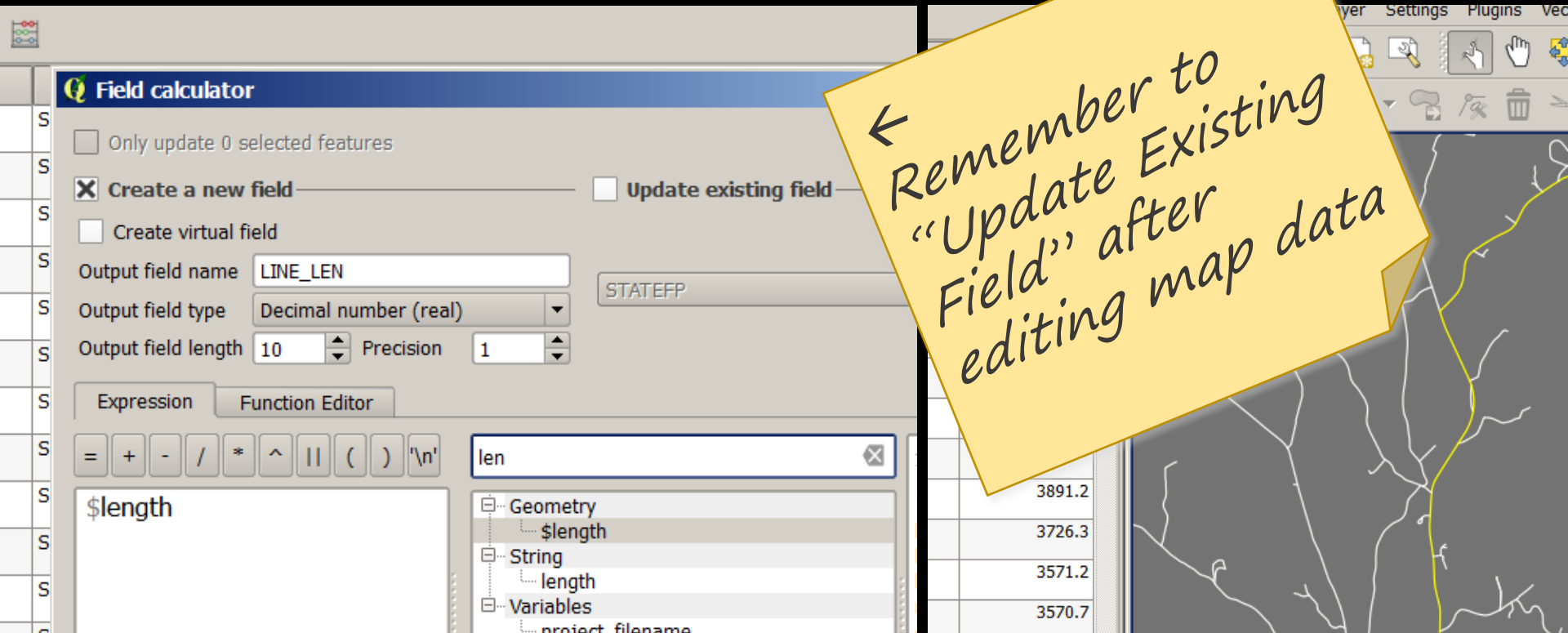
METHOD: Field Calculator



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

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

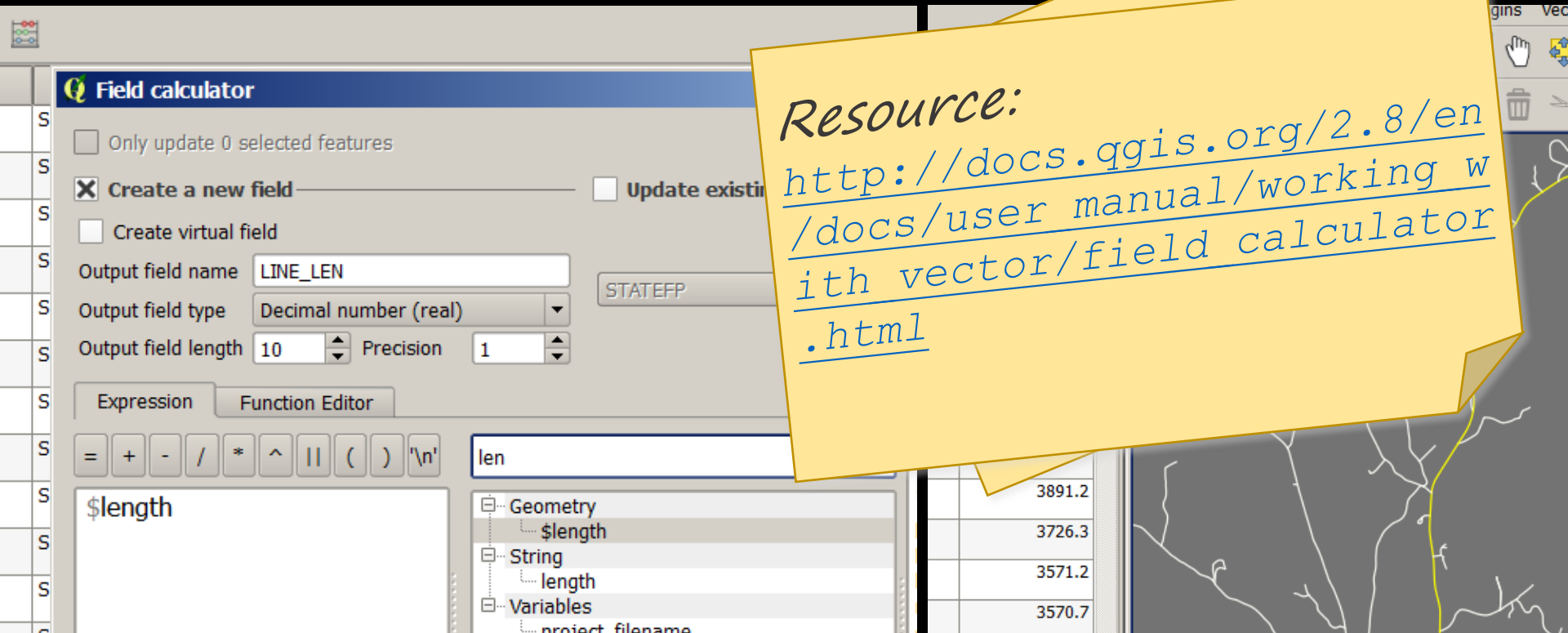
METHOD: Field Calculator



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



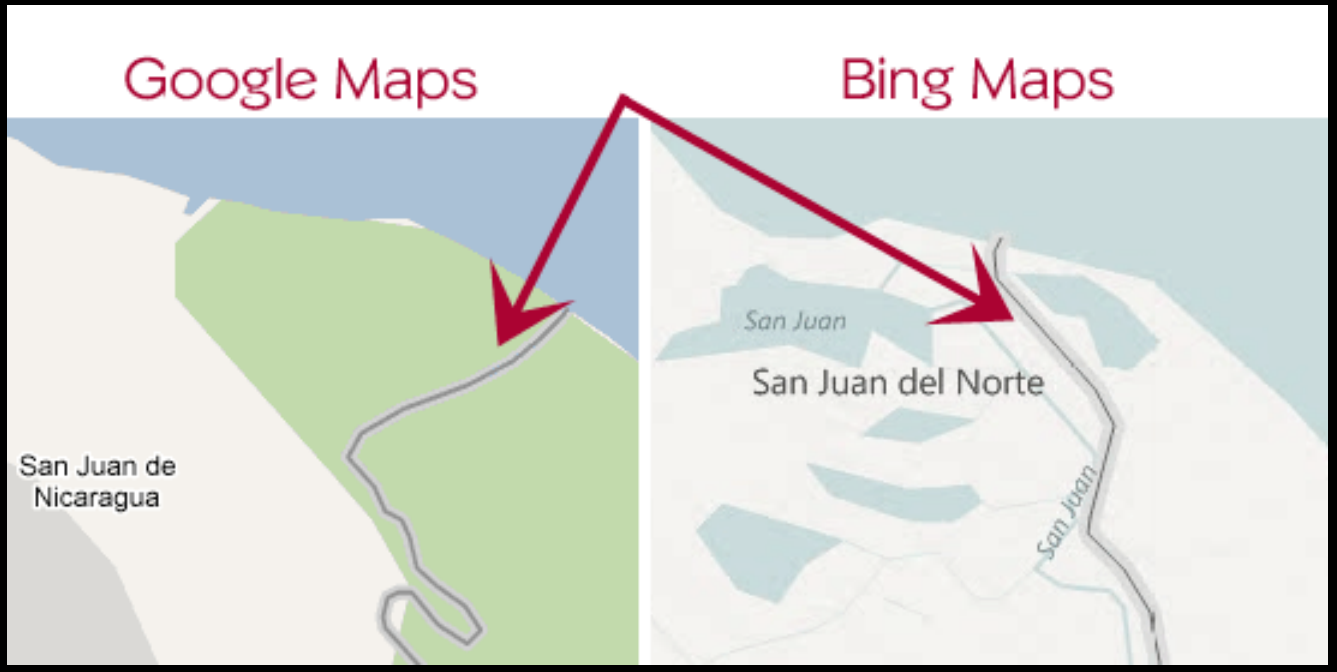
Resource:
http://docs.qgis.org/2.8/en/docs/user_manual/working_with_vector/field_calculator.html

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.

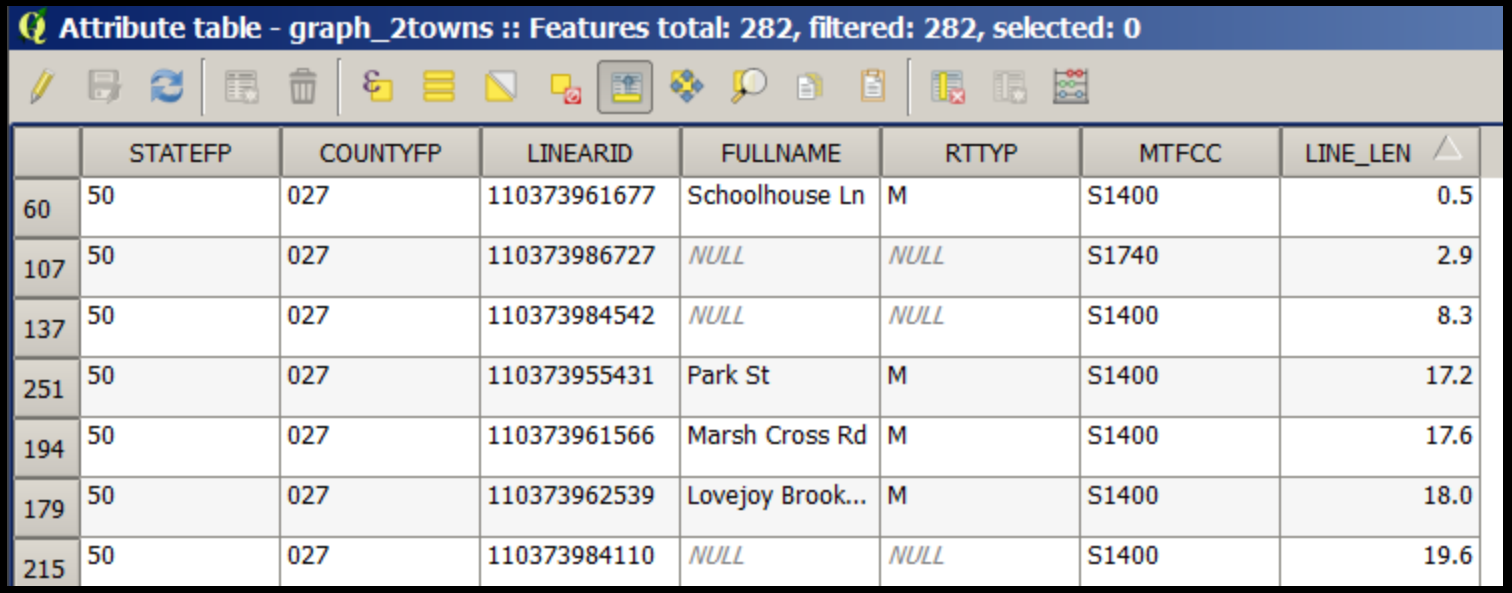
2. Specs and Checks

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: Calcu Length

Not scalable

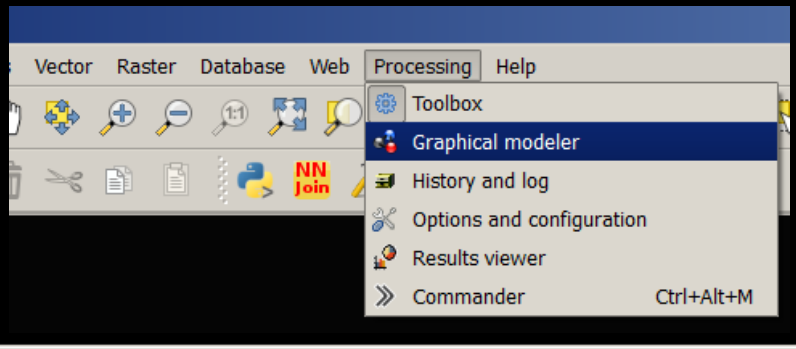
Attribute table - graph_2towns ::

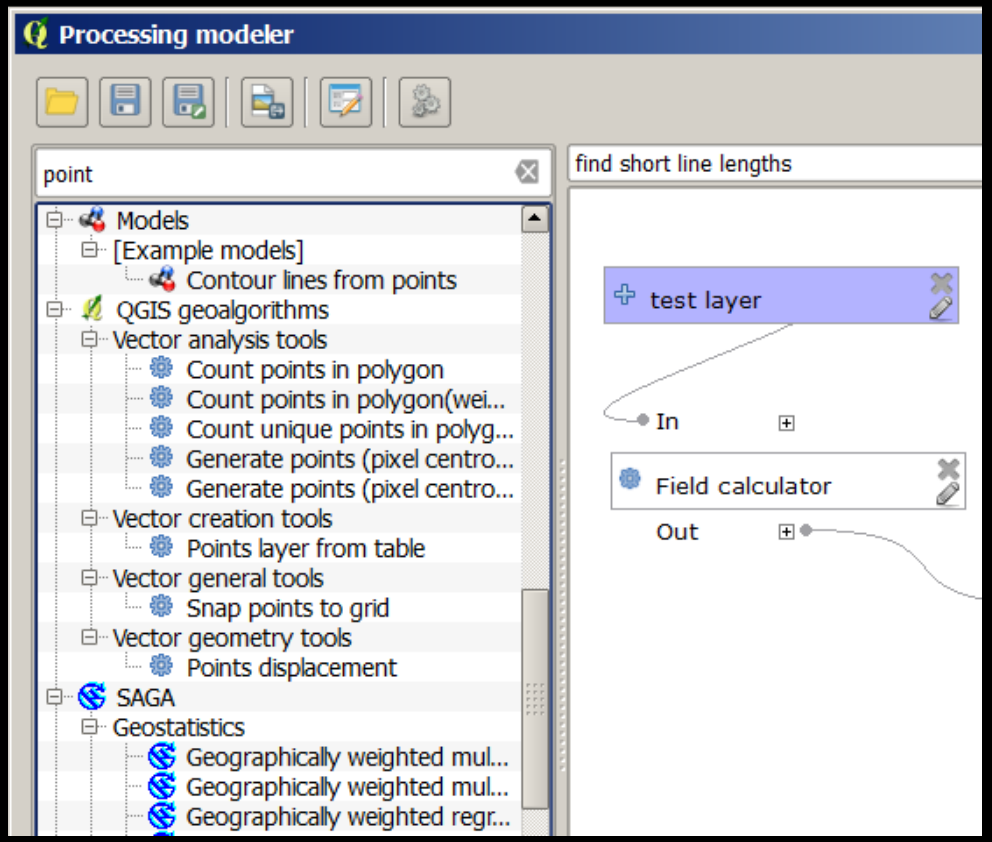
	STATEFP	COUNTYFP				TFCC	LINE_LEN
60	50	027	110			00	0.5
107	50	027	1103			S1740	2.9
137	50	027	110373984542	NULL	NULL	S1400	8.3
251	50	027	110373955431	Park St	M	S1400	17.2
194	50	027	110373961566	Marsh Cross Rd	M	S1400	17.6
179	50	027	110373962539	Lovejoy Brook...	M	S1400	18.0
215	50	027	110373984110	NULL	NULL	S1400	19.6

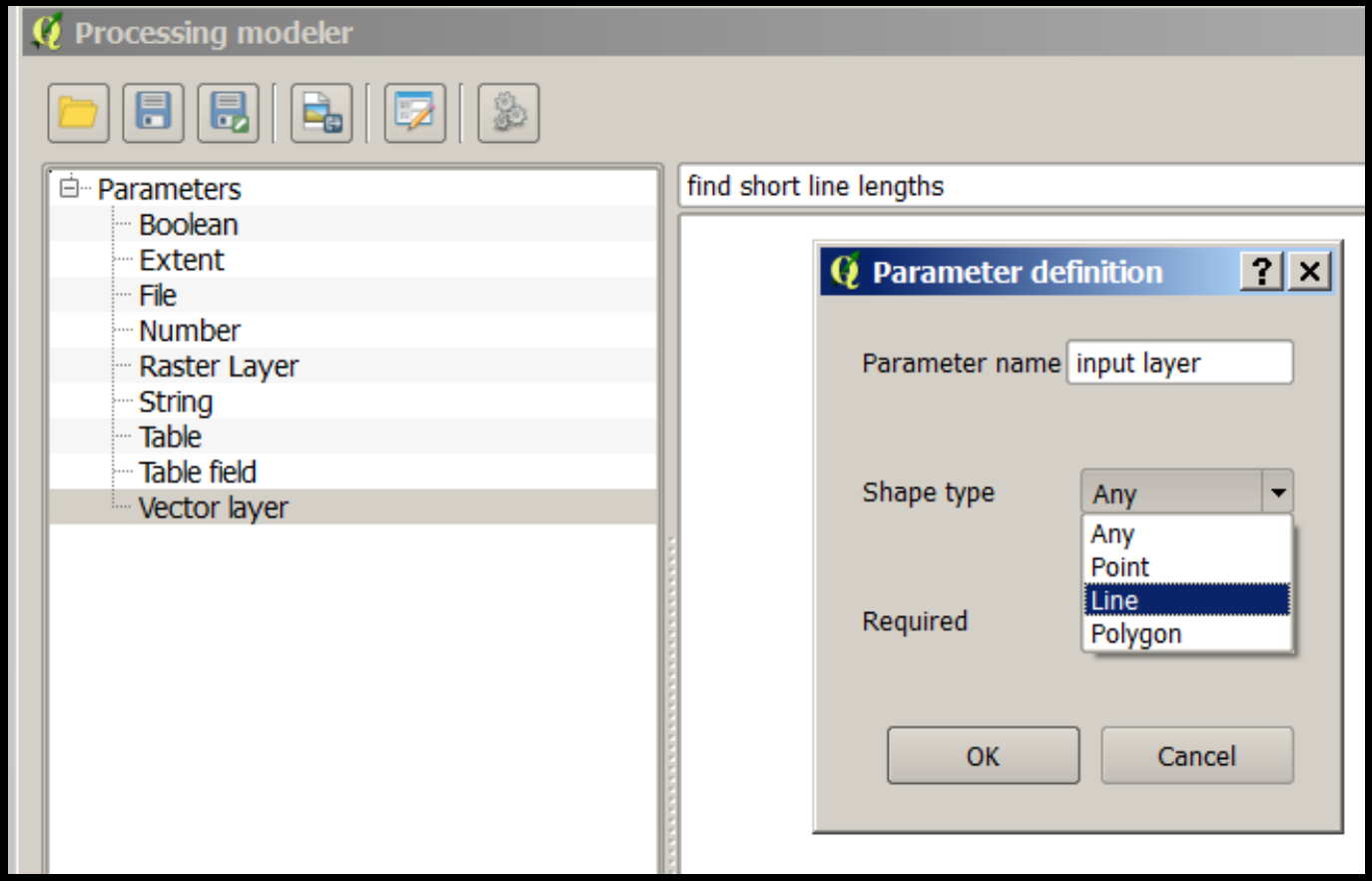
D. Captured roads exist in reality.

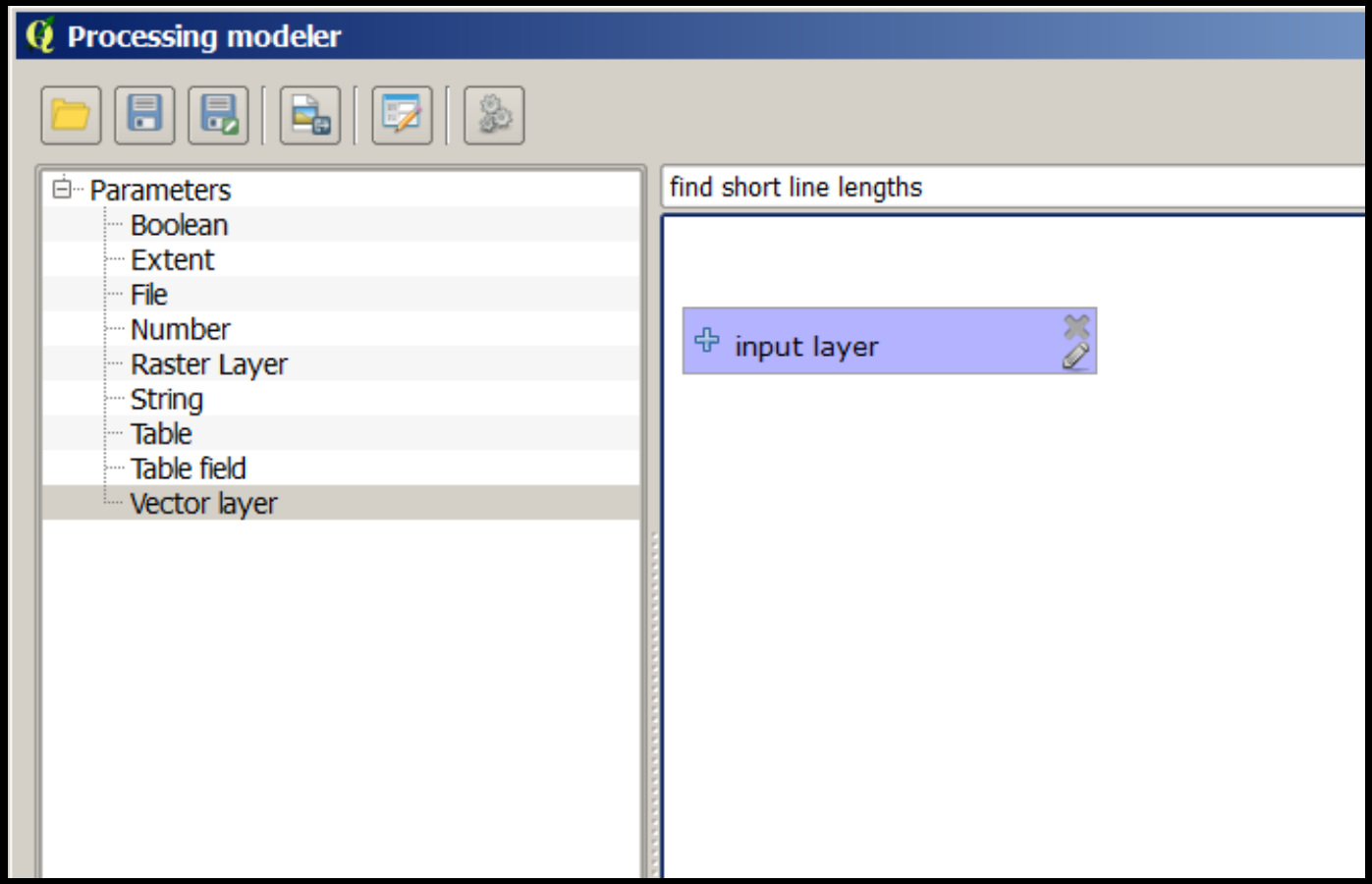
METHOD 2: Model

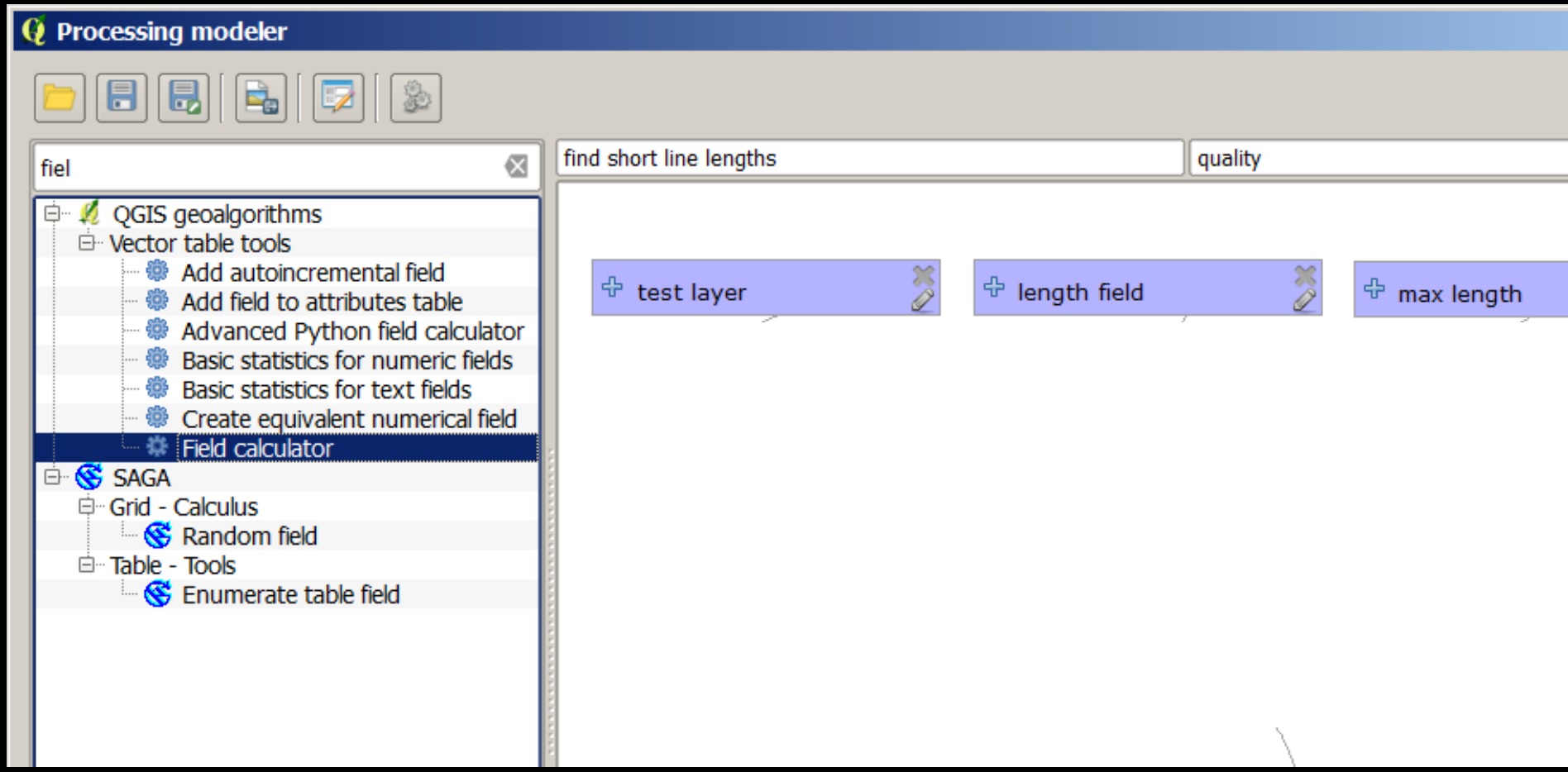
- Create multi-step workflow
- Graphical interface
- Powerful processing tools
- Similar to ArcMap ModelBuilder FME Workbench

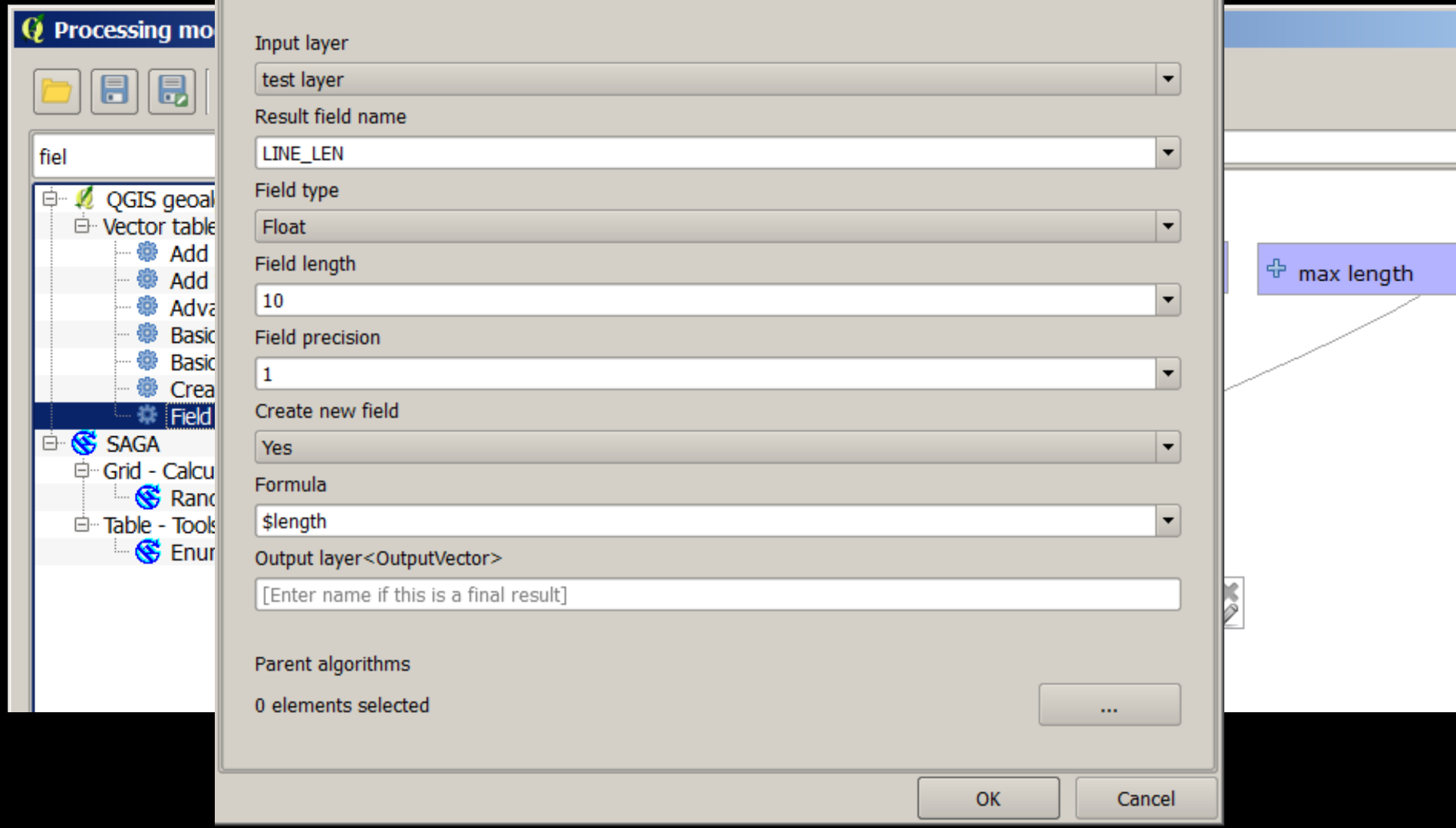


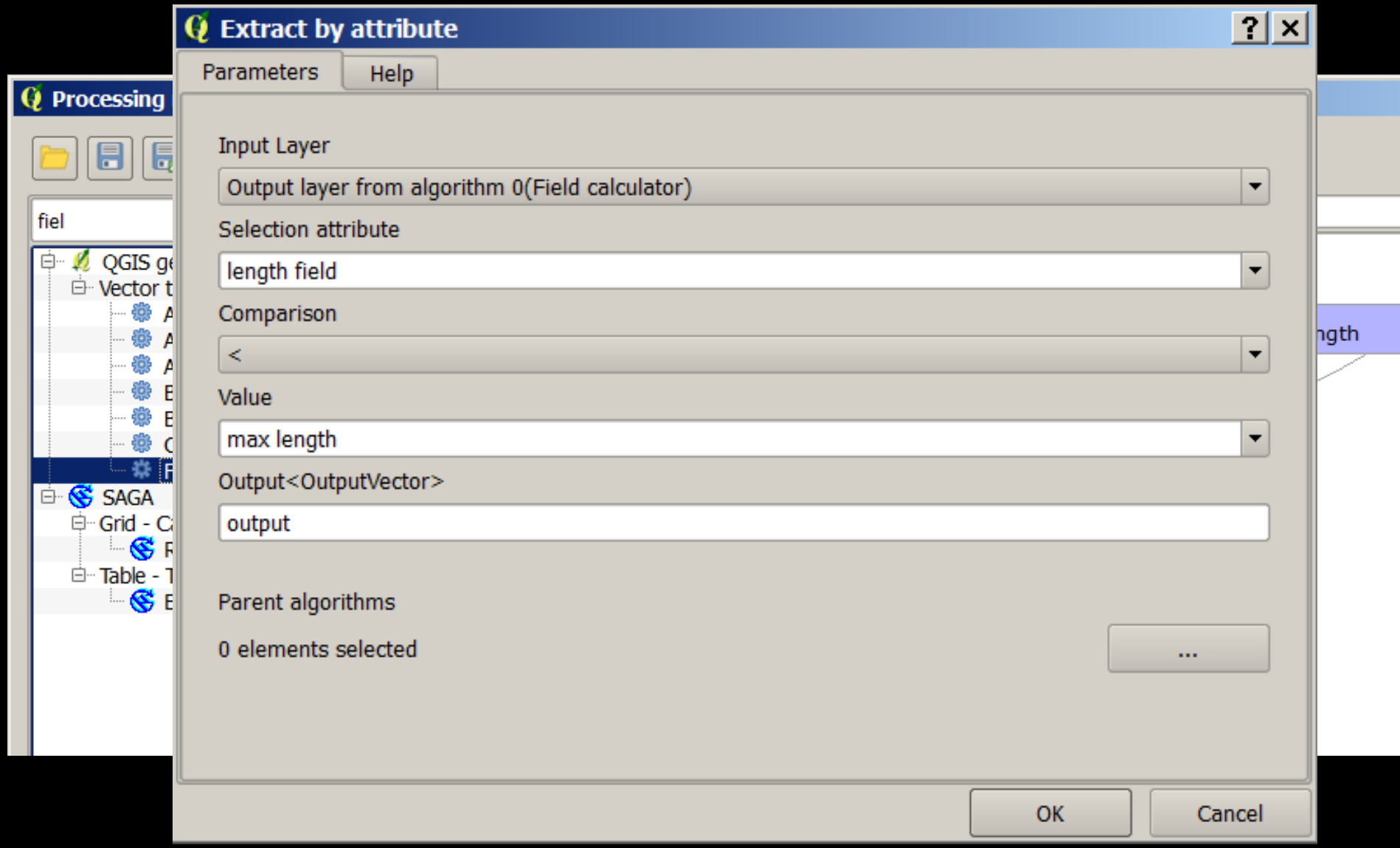


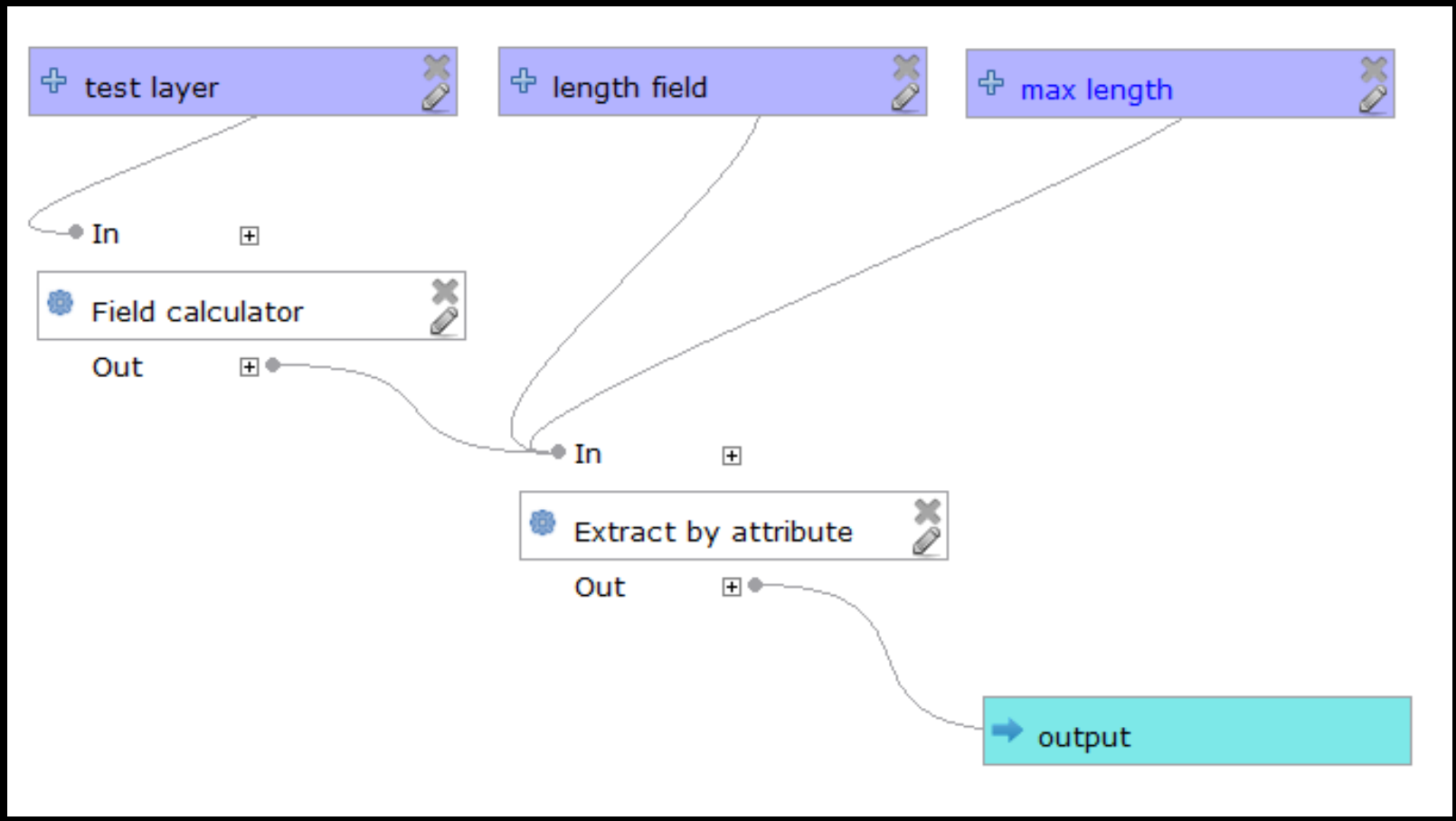

Processing modeler
Parameters
Boolean
Extent
File
Number
Raster Layer
String
Table
Table field
Vector layer
find short line lengths
Parameter definition
Parameter name
input layer
Shape type
Any
Any
Point
Line
Polygon
Required
OK
Cancel

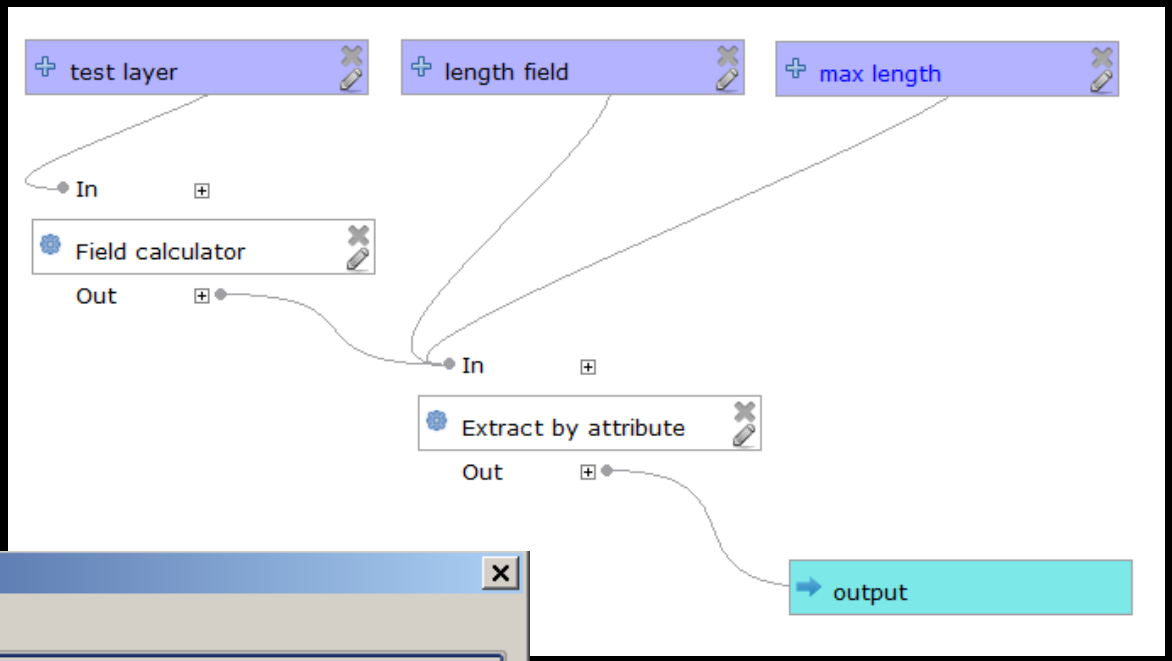

Processing modeler
Parameters
Boolean
Extent
File
Number
Raster Layer
String
Table
Table field
Vector layer
find short line lengths
input layer


Processing modeler
fiel
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
find short line lengths
quality
test layer
length field
max length


Processing mo
fiel
QGIS geoal
Vector table
Add
Add
Adva
Basic
Basic
Crea
Field
SAGA
Grid - Calcu
Ranc
Table - Tools
Enur
max length
Field calculator
Parameters
Help
Input layer
test layer
Result field name
LINE_LEN
Field type
Float
Field length
10
Field precision
1
Create new field
Yes
Formula
$length
Output layer<OutputVector>
[Enter name if this is a final result]
Parent algorithms
0 elements selected
...
OK
Cancel


Processing
fiel
QGIS ge
Vector t
SAGA
Grid - C
Table - T
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


test layer
length field
max length
In
Field calculator
Out
In
Extract by attribute
Out
output



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 ...

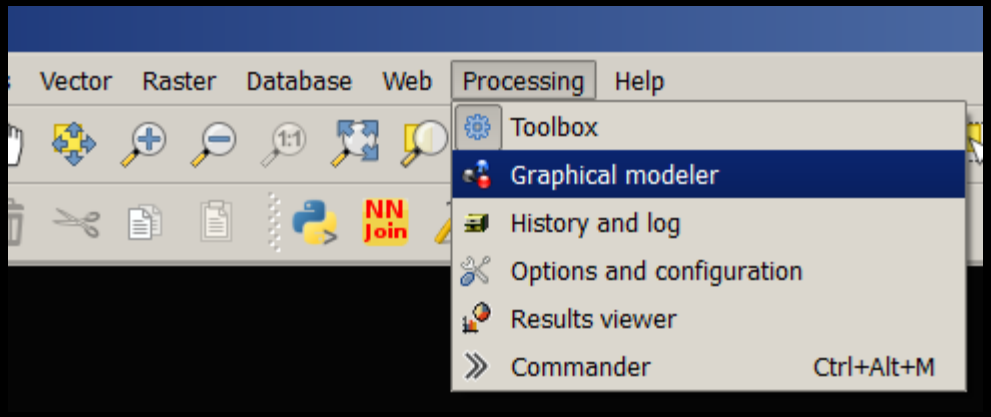
Open output file after running algorithm

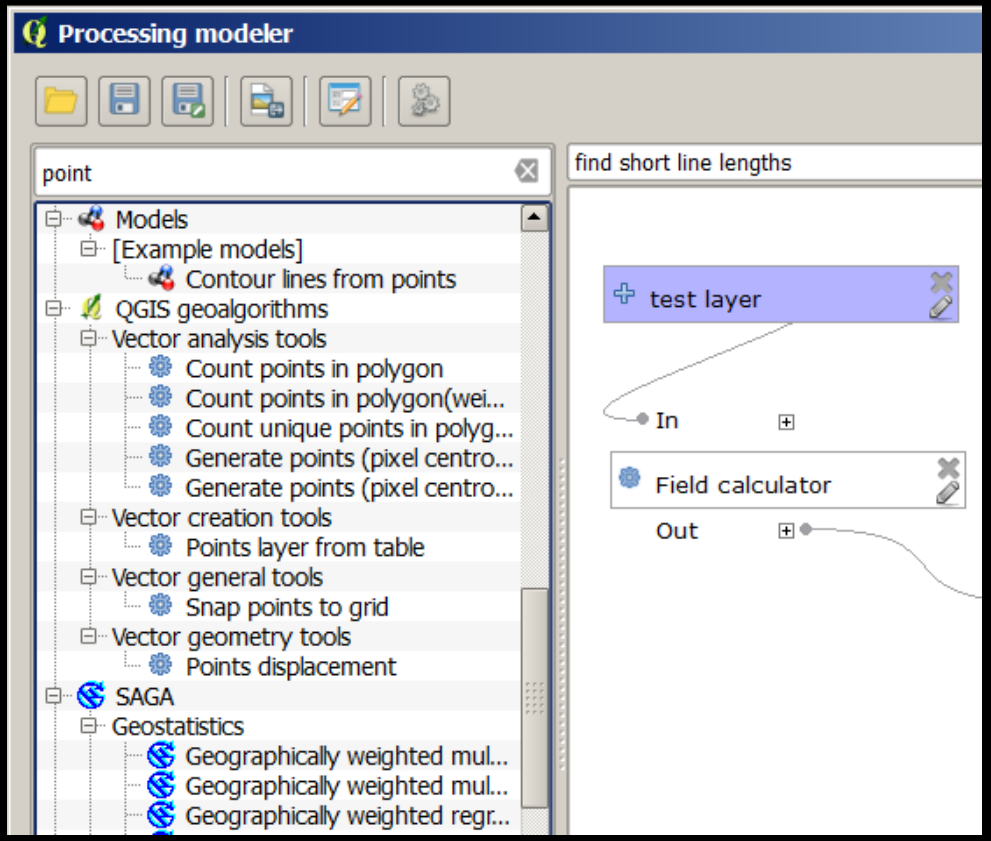
0%

Run | Close | Cancel

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.

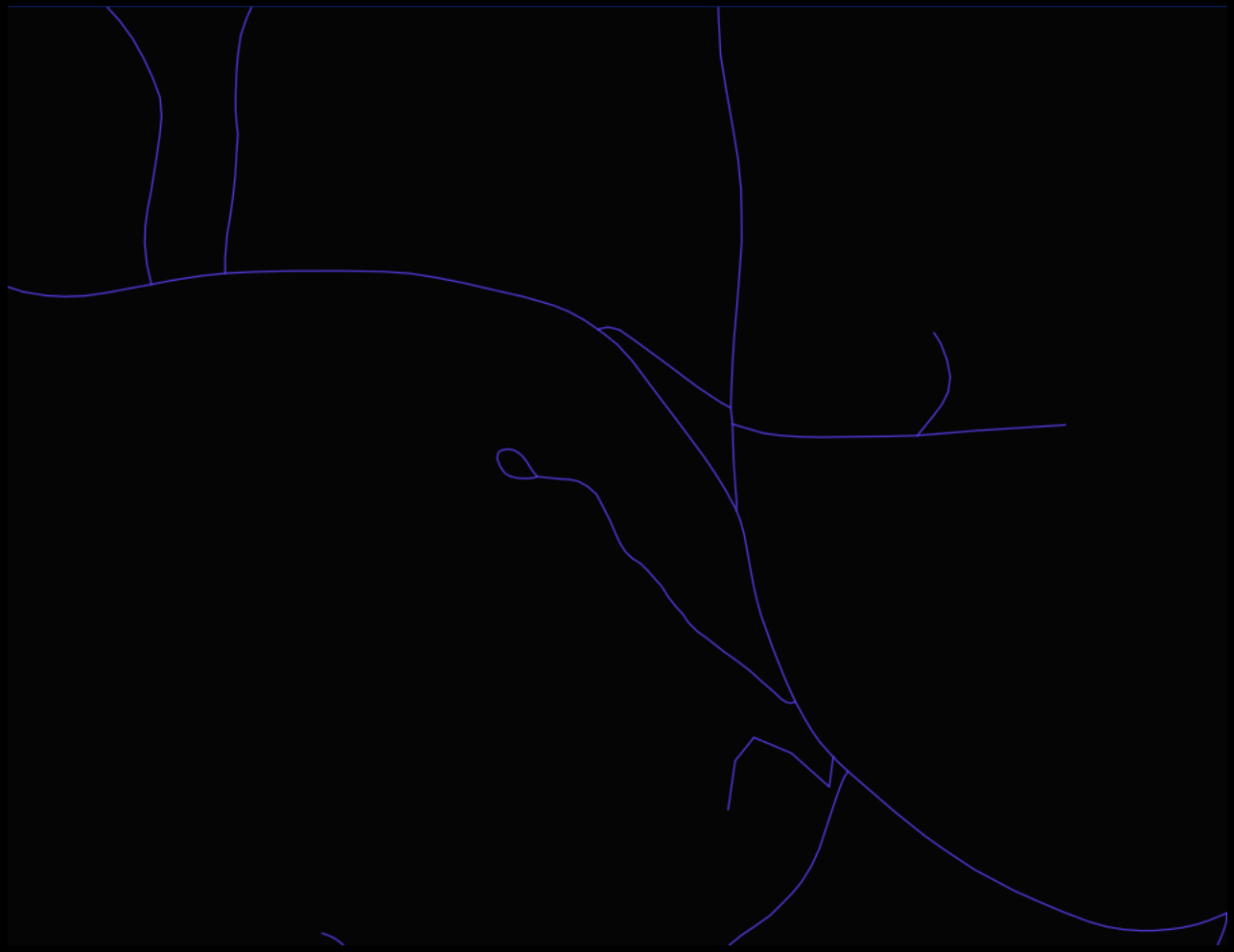
CHECK: For a Road, the entire feature must be captured within 10m of a feature in the designated reference dataset

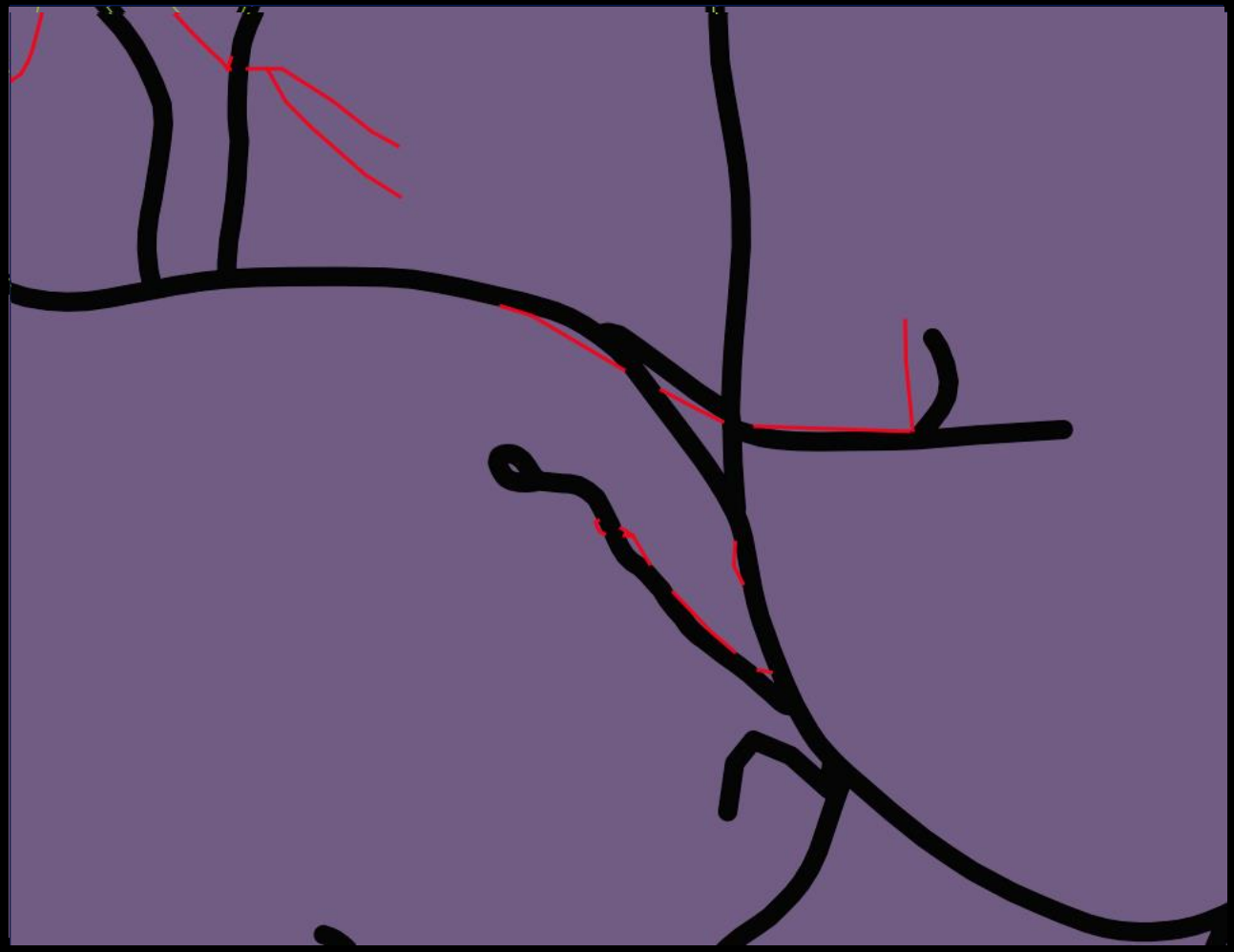
METHOD: Buffer

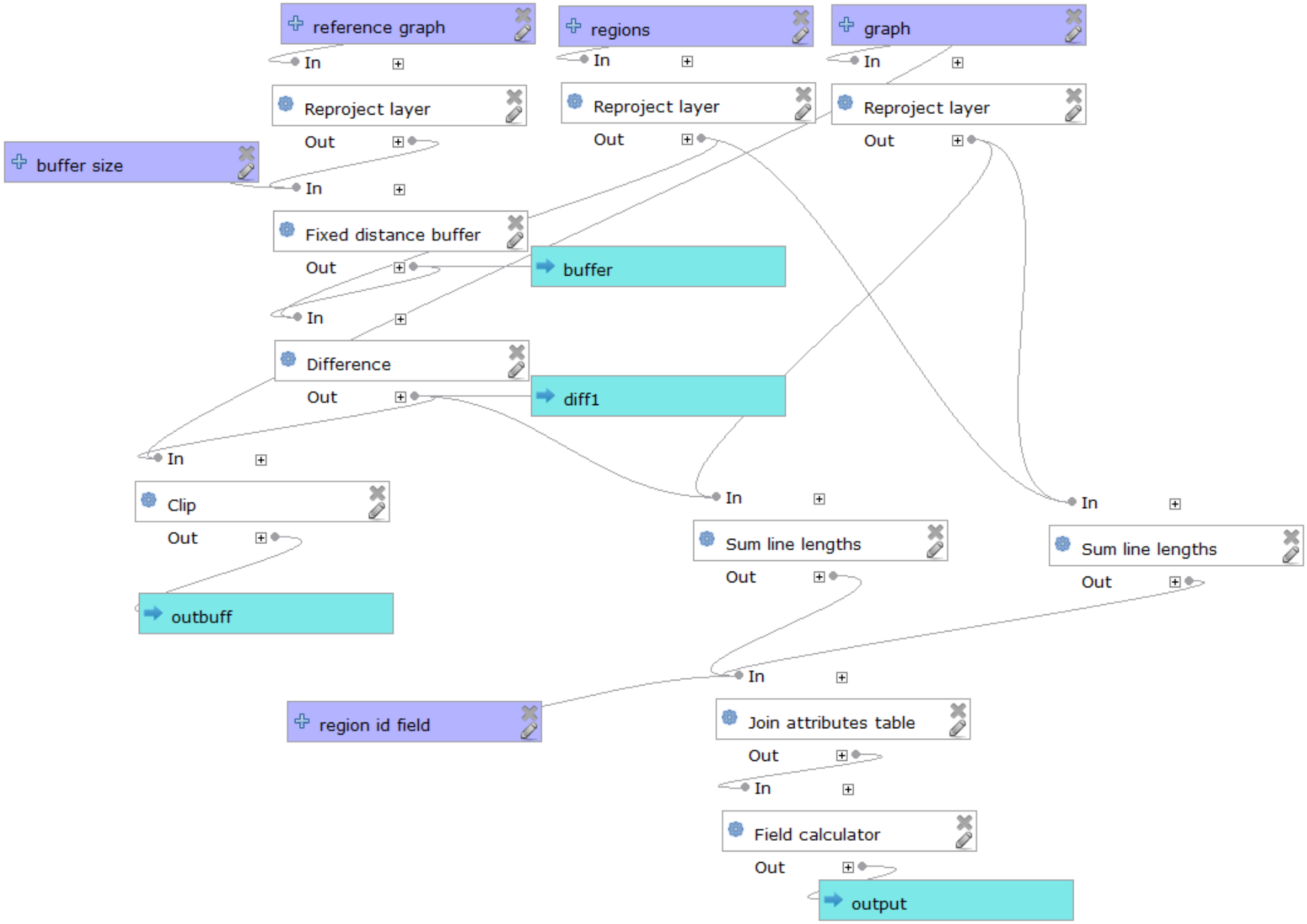
D. Captured roads exist in reality.

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

METHOD: Buffe



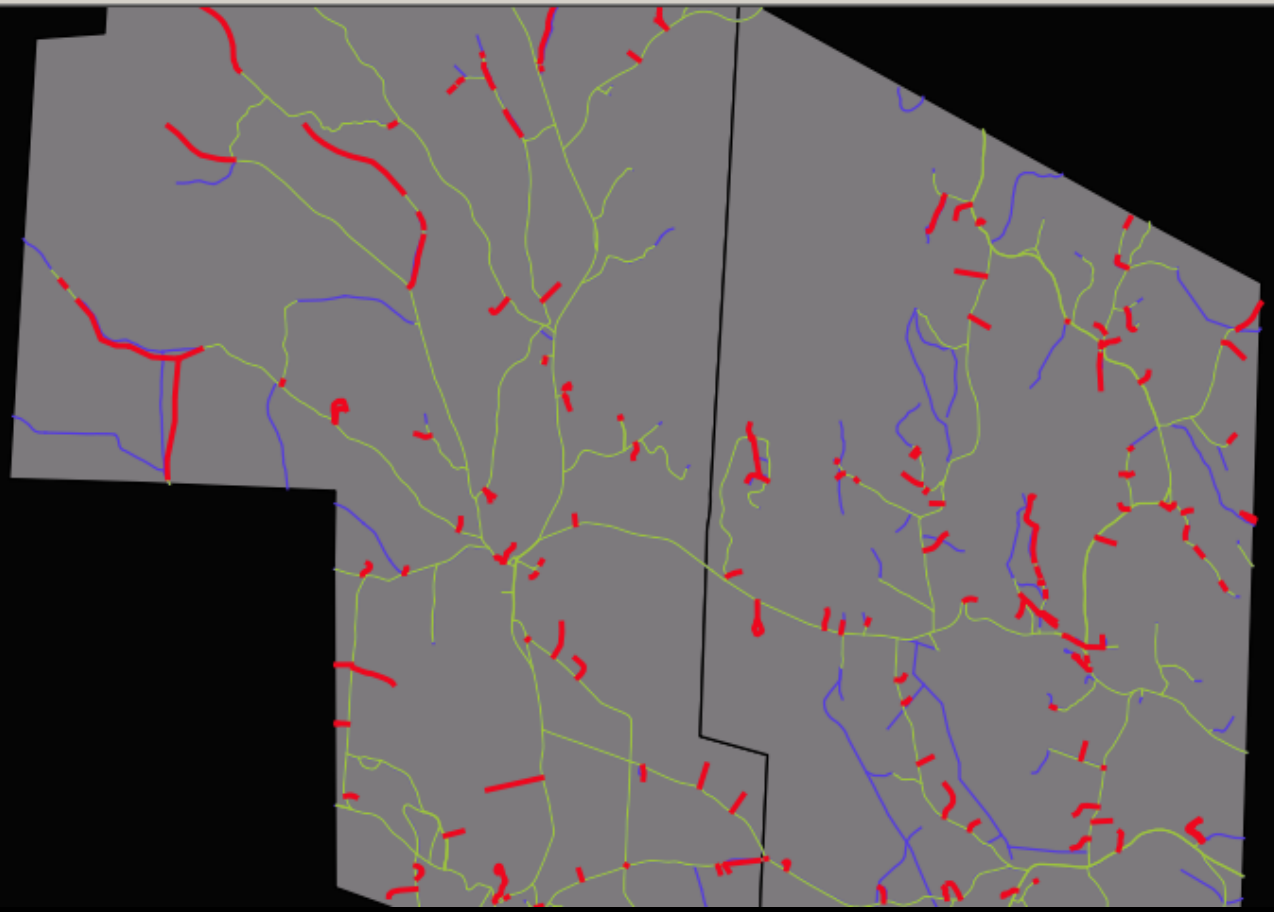



reference graph
regions
graph
In
In
In
Reproject layer
Reproject layer
Reproject layer
Out
Out
Out
buffer size
In
Fixed distance buffer
Out
buffer
In
Difference
Out
diff1
In
In
In
Clip
Out
Sum line lengths
Out
Sum line lengths
Out
outbuff
In
region id field
Join attributes table
Out
In
Field calculator
Out
output

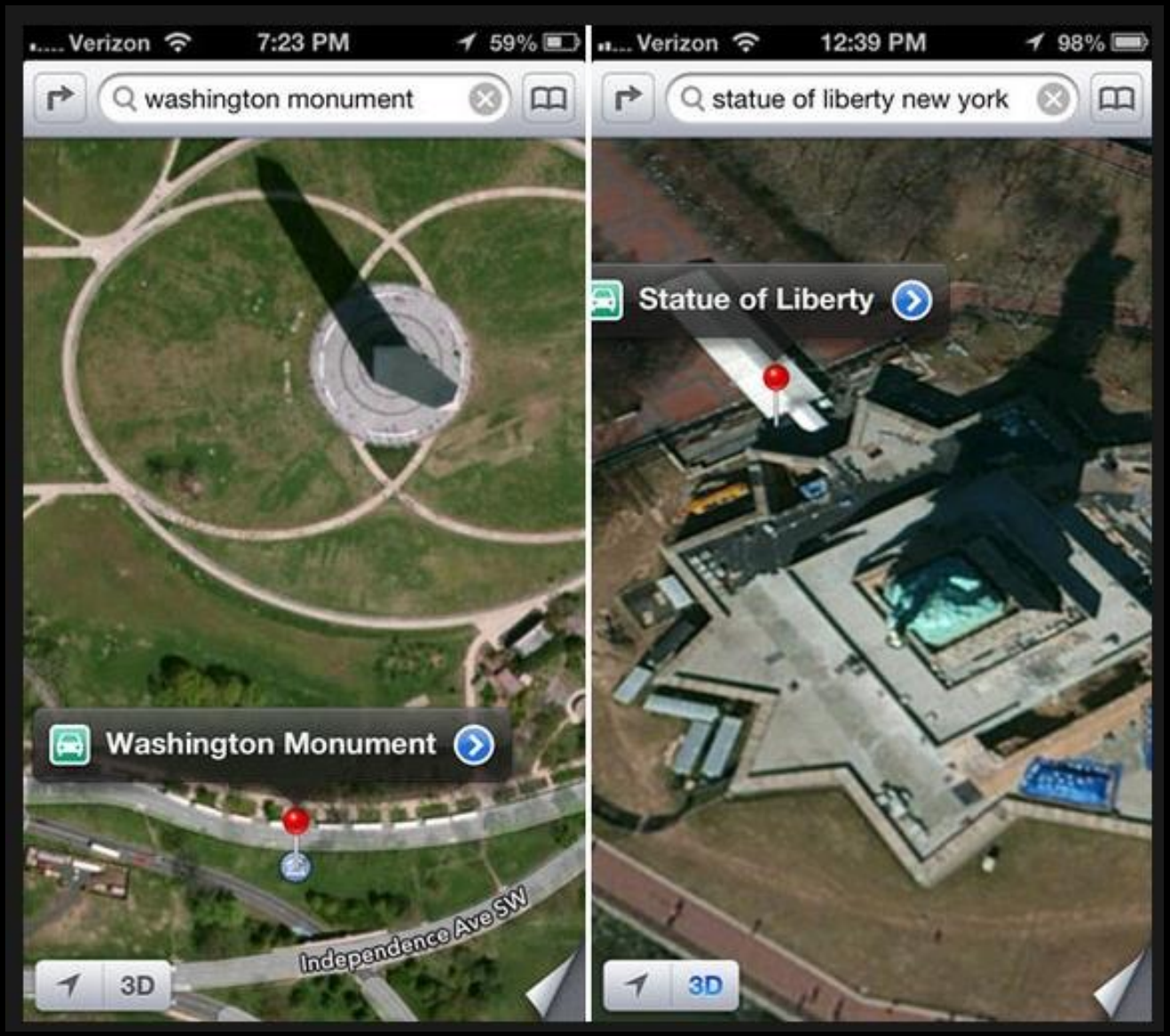
Attribute table - output :: Features total: 2, filtered: 2, selected: 0

	FIPS6	TOWNNAME	totalLEN	totalCNT	fips6_2	townname_2	outbufLEN	outbufCNT	outbufPERC
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...	102.00000000...	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.

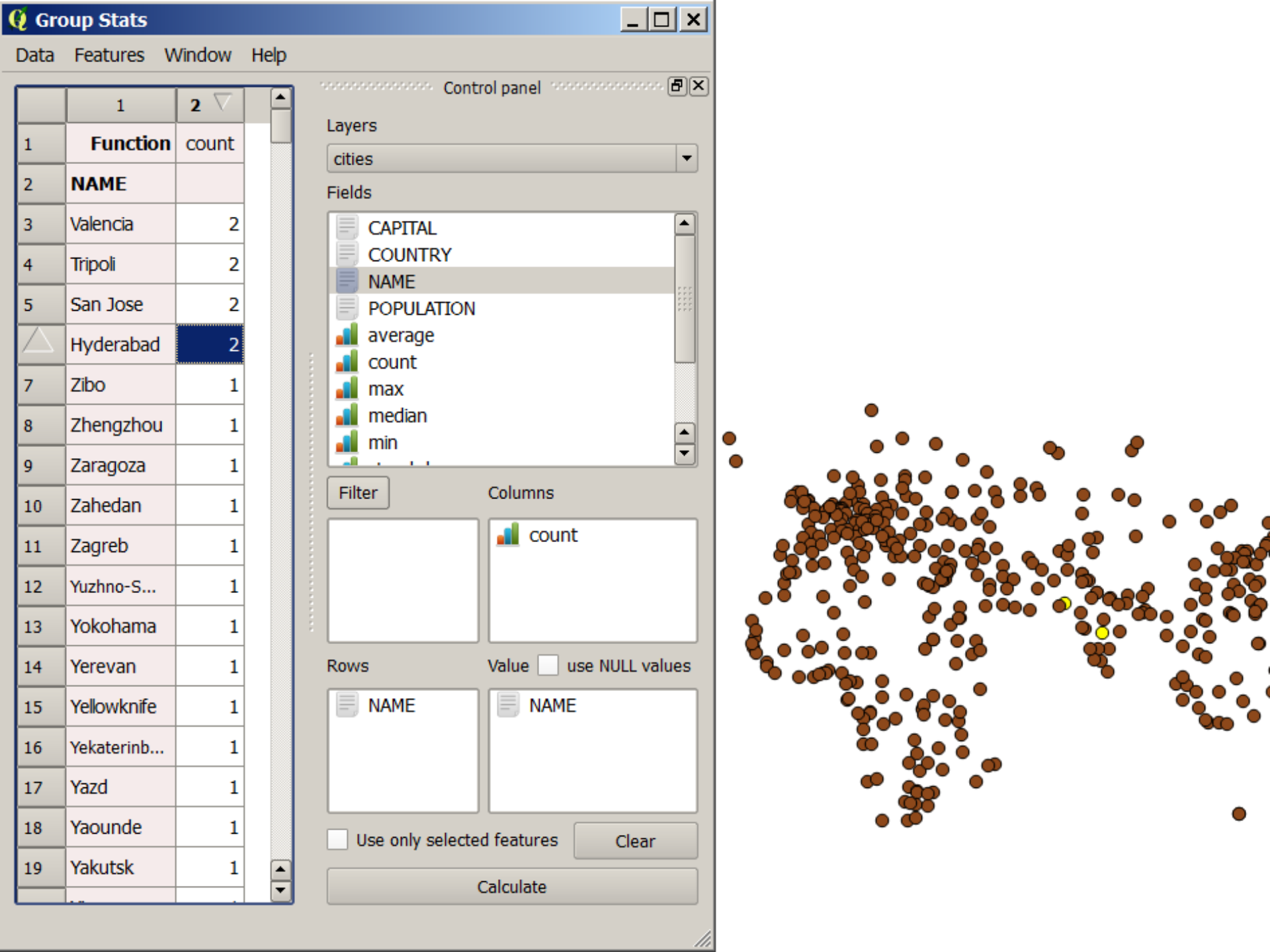
2. Specs and Checks

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

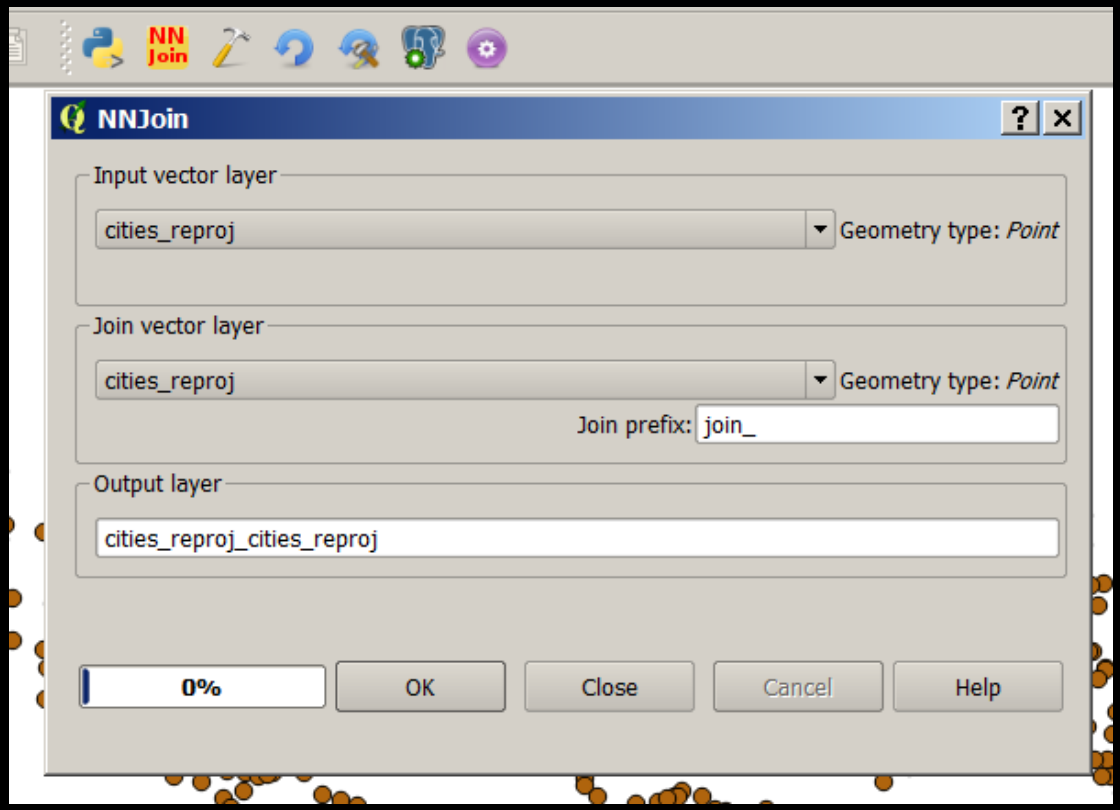


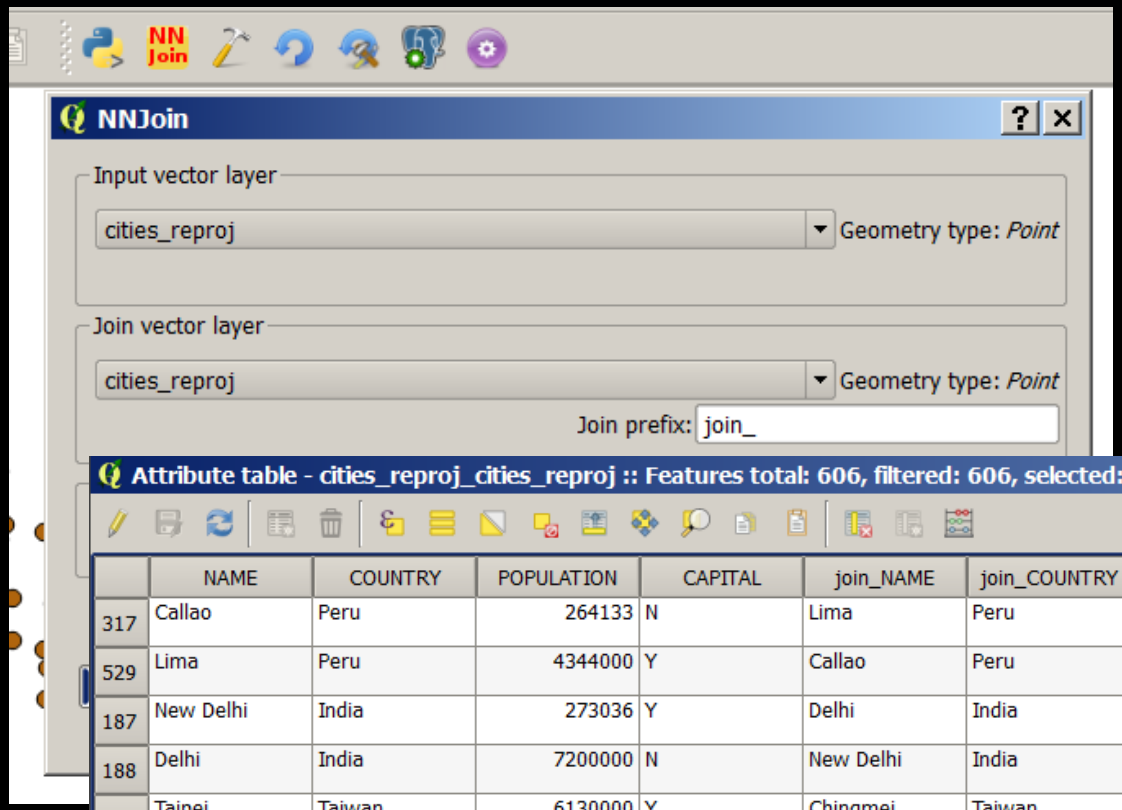
	1	2
1	**Function**	count
2	**NAME**	
3	Valencia	2
4	Tripoli	2
5	San Jose	2
	Hyderabad	2
7	Zibo	1
8	Zhengzhou	1
9	Zaragoza	1
10	Zahedan	1
11	Zagreb	1
12	Yuzhno-S...	1
13	Yokohama	1
14	Yerevan	1
15	Yellowknife	1
16	Yekaterinb...	1
17	Yazd	1
18	Yaounde	1
19	Yakutsk	1

E. Features are not duplicated.

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

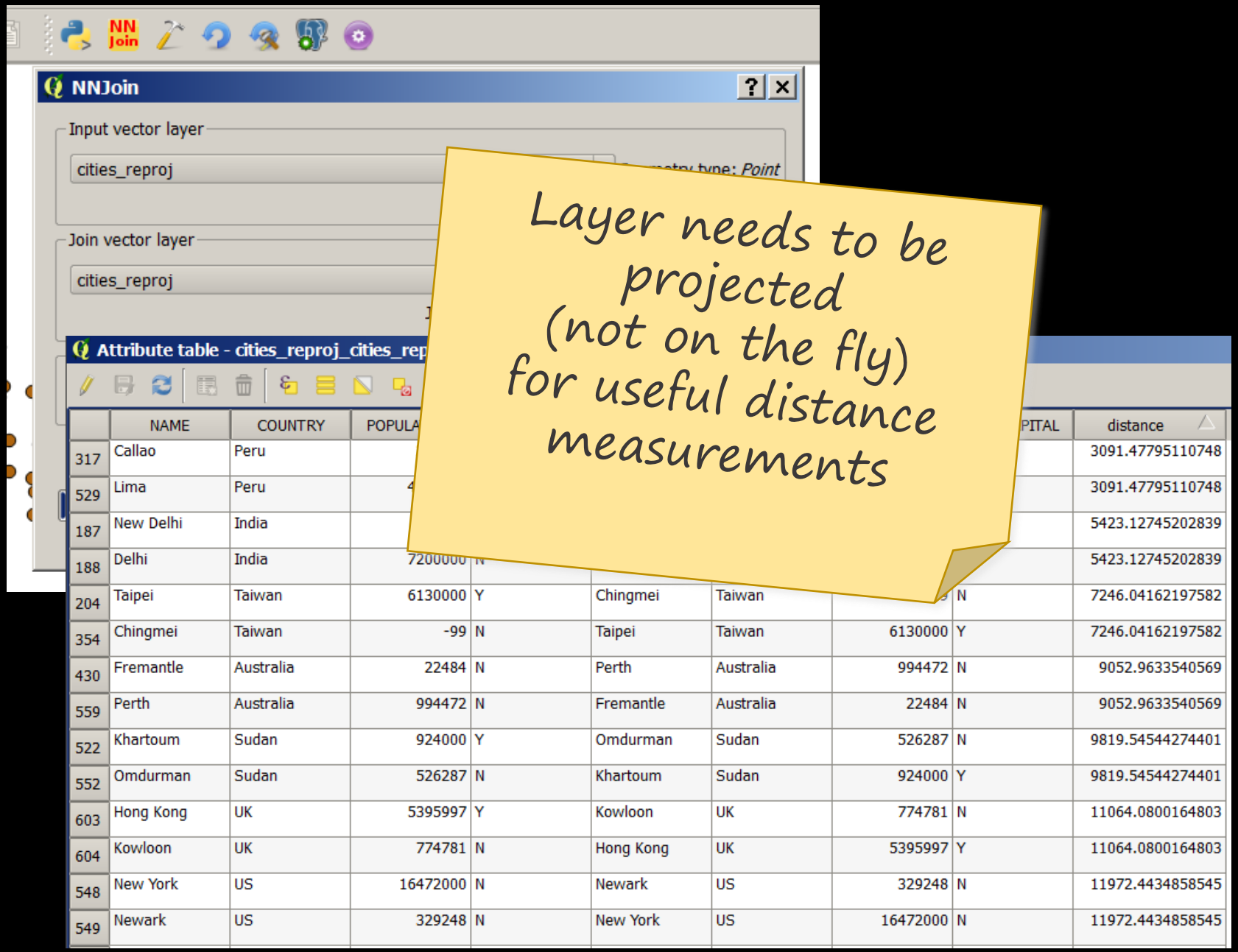
METHOD: Plugin: NNJoin


NN Join
NNJoin
Input vector layer
cities_reproj
Geometry type: Point
Join vector layer
cities_reproj
Geometry type: Point
Join prefix: join_
Output layer
cities_reproj_cities_reproj
0%
OK
Close
Cancel
Help



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

	NAME	COUNTRY	POPULATION	CAPITAL	join_NAME	join_COUNTRY	ɔin_POPULATIOI	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



	NAME	COUNTRY	POPULA				PITAL	distance
317	Callao	Peru						3091.47795110748
529	Lima	Peru	4					3091.47795110748
187	New Delhi	India						5423.12745202839
188	Delhi	India	7200000	N				5423.12745202839
204	Taipei	Taiwan	6130000	Y	Chingmei	Taiwan	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

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.


Schrade Rd
7601 Cousteau Dr
Supposed to be demolished.
Cousteau Dr
Cousteau Dr
Actually demolished.
7601 Calypso Dr
(Wrong address still.)
7601 Cousteau Dr
Rowlett, TX 75088
32.899449, -96.527944
Street View
Calypso Dr


COUSTEAU
DR
4200

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?***

http://tinyurl.com/qgis-quality-0516