Hotspot Analysis with QGIS

Demo Exercise

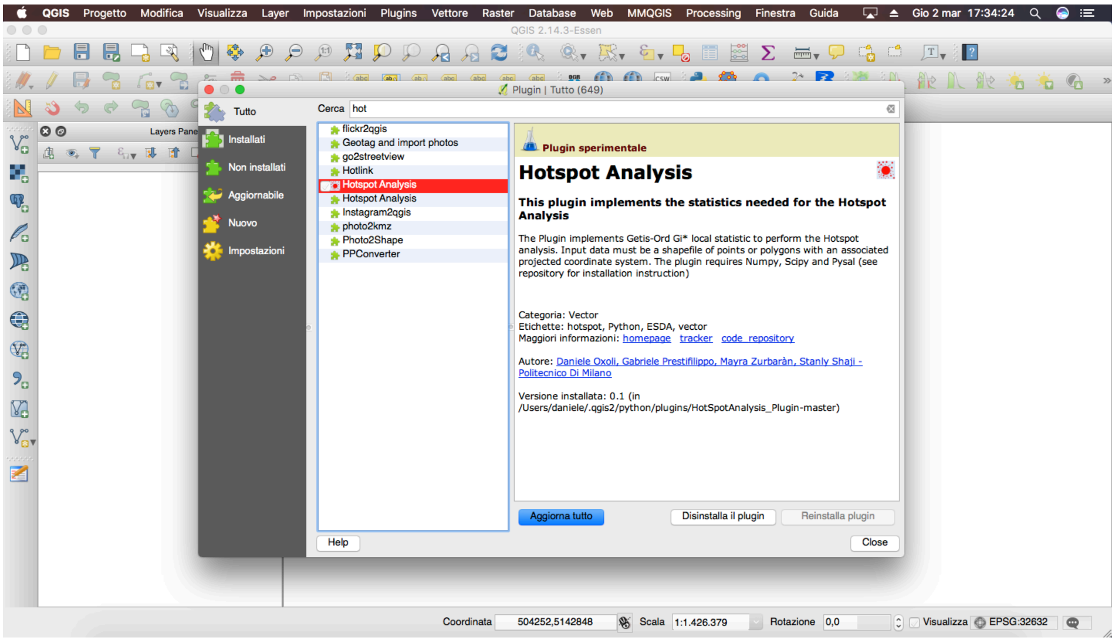
The following demo includes two examples of the QGIS Hotspot Analysis plugin application.

Plugin update: 10/03/2017

Content:

0. System check

After the installation of both the plugin and plugin's dependencies:
(see: https://github.com/danioxoli/HotSpotAnalysis_Plugin/README.md)
Open QGIS and check if the plugin is correctly installed:



Plugin sperimentale

Hotspot Analysis

This plugin implements the statistics needed for the Hotspot Analysis

The Plugin implements Getis-Ord Gi* local statistic to perform the Hotspot analysis. Input data must be a shapefile of points or polygons with an associated projected coordinate system. The plugin requires Numpy, Scipy and Pysal (see repository for installation instruction)

Categoria: Vector
Etichette: hotspot, Python, ESDA, vector
Maggiori informazioni: homepage tracker code_repository

Autore: Daniele Oxoli, Gabriele Prestifilippo, Mayra Zurbaràn, Stanly Shaji - Politecnico Di Milano

Versione installata: 0.1 (in /Users/daniele/.qgis2/python/plugins/HotSpotAnalysis_Plugin-master)

No Python errors should appear while opening QGIS or installing the plugin. If this is not true, please check that the installation procedure was performed correctly.
(Sometimes, due to the system Python Path some unexpected error may occur. If these are not solved by repeating the installation procedure of both plugin and dependencies, please report the issue to the authors)

1. Create a valid input layer

The plugin accepts as input points or polygons shapefile with at least a numeric attribute assigned to any geometry of the dataset.

1.1 Points Shapefile

A demo input points shapefile is available here:
https://github.com/danioxoli/HotSpotAnalysis_Plugin/tree/master/test_data - > **demo_data**
Open the **demo_data** shapefile and look at its attribute table:

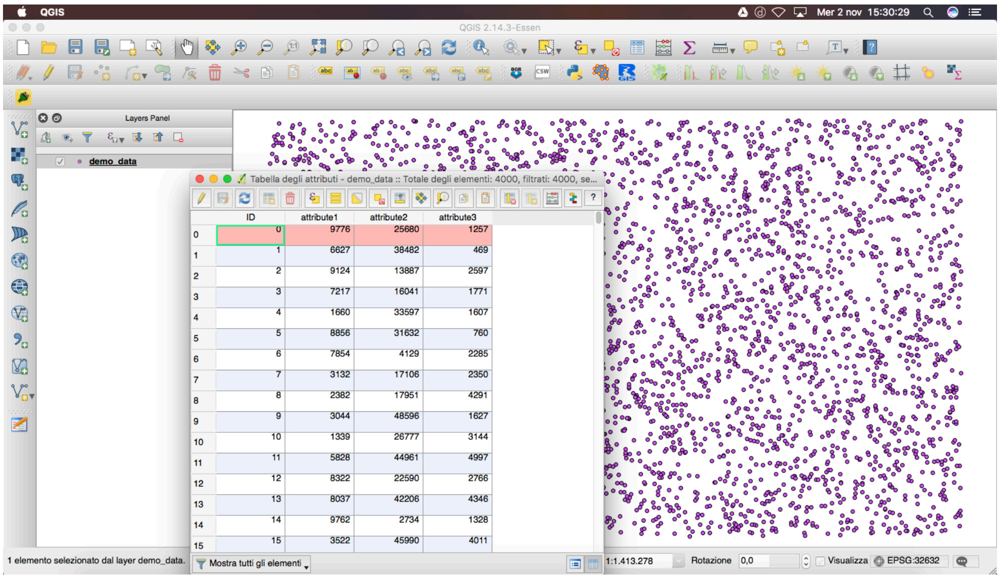


Tabella degli attributi - demo_data :: Totale degli elementi: 4000, filtrati: 4000, se...

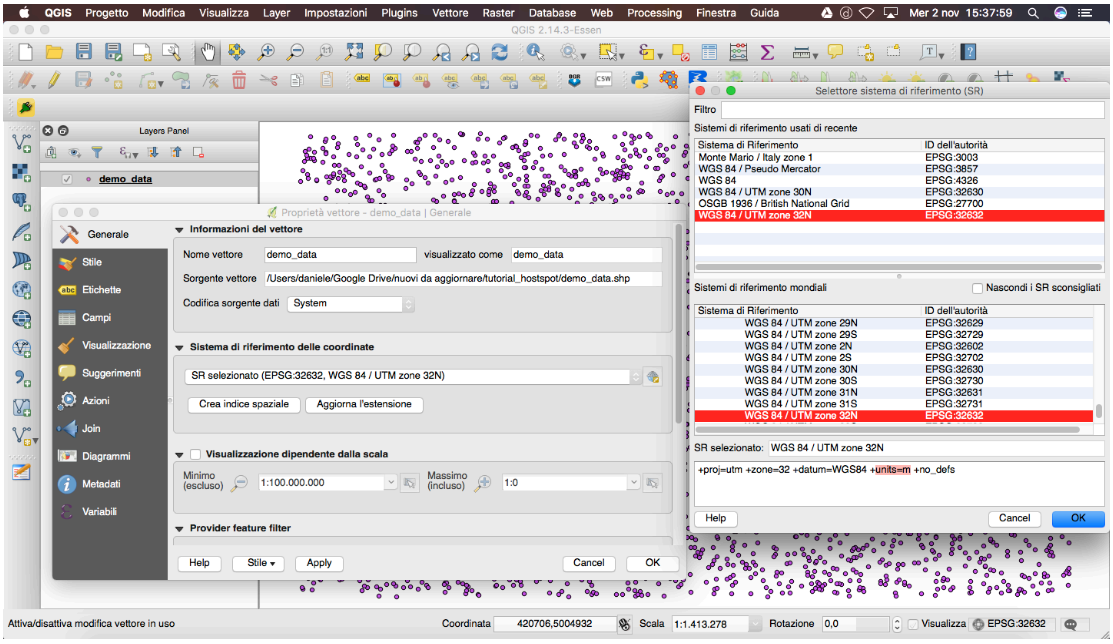
	ID	attribute1	attribute2	attribute3
0	0	9776	25680	1257
1	1	6627	38482	469
2	2	9124	13887	2597
3	3	7217	16041	1771
4	4	1660	33597	1607
5	5	8856	31632	760
6	6	7854	4129	2285
7	7	3132	17106	2350
8	8	2382	17951	4291
9	9	3044	48596	1627
10	10	1339	26777	3144
11	11	5828	44961	4997
12	12	8322	22590	2766
13	13	8037	42206	4346
14	14	9762	2734	1328
15	15	3522	45990	4011

1 elemento selezionato dal layer demo_data.

As it can be seen, to any of the 4000 points of **demo_data** are assigned three numeric attributes. Your dataset should look like this latter.

To properly run Hotspot Analysis your dataset should contain at least ~ 30 points (indicatively).
Another requirement is the coordinate system associated to your input shapefile. The plugin requires your layer to be **projected**. Therefore, be sure about that by checking at the **layer properties**. Moreover, be aware about the **unit of measure** in which your selected projected coordinate system is expressed. With respect to the **demo_data** (see picture below), the assigned projected coordinate system is WGS84/UTM32N which is expressed in **meters**. The unit of measure is of primary importance to initialise the plugin running parameters (i.e. Distance Band).

The plugin implements the Getis-Ord Gi local statistic, which aims to detected atypical locations (i.e. hotspots/coldspots) in the spatial arrangement of a given variable. Practically speaking, Gi* compares local averages with global average to underline the presence of significant high-values (or low-values)* *clusters. Local averages are computed by considering, for any location of the dataset, a set of neighbour elements within a specific distance from the focal position. For this reason, the plugin requires to specify a Distance Band by using the same unit of measure of the projected coordinate system of the input shapefile.*



Proprietà vettore - demo_data | Generale

Informazioni del vettore

Nome vettore: demo_data visualizzato come: demo_data

Sorgente vettore: /Users/daniele/Google Drive/nuovi da aggiornare/tutorial_hostspot/demo_data.shp

Codifica sorgente dati: System

Sistema di riferimento delle coordinate

SR selezionato (EPSG:32632, WGS 84 / UTM zone 32N)

□ Visualizzazione dipendente dalla scala

Minimo (escluso): 1:100.000.000 Massimo (incluso): 1:0

Provider feature filter

Selettore sistema di riferimento (SR)

Filtro

Sistemi di riferimento usati di recente

Sistema di Riferimento	ID dell'autorità
Monte Mario / Italy zone 1	EPSG:3003
WGS 84 / Pseudo Mercator	EPSG:3857
WGS 84	EPSG:4326
WGS 84 / UTM zone 30N	EPSG:32630
OSGB 1936 / British National Grid	EPSG:27700
WGS 84 / UTM zone 32N	EPSG:32632

Sistemi di riferimento mondiali □ Nascondi i SR sconsigliati

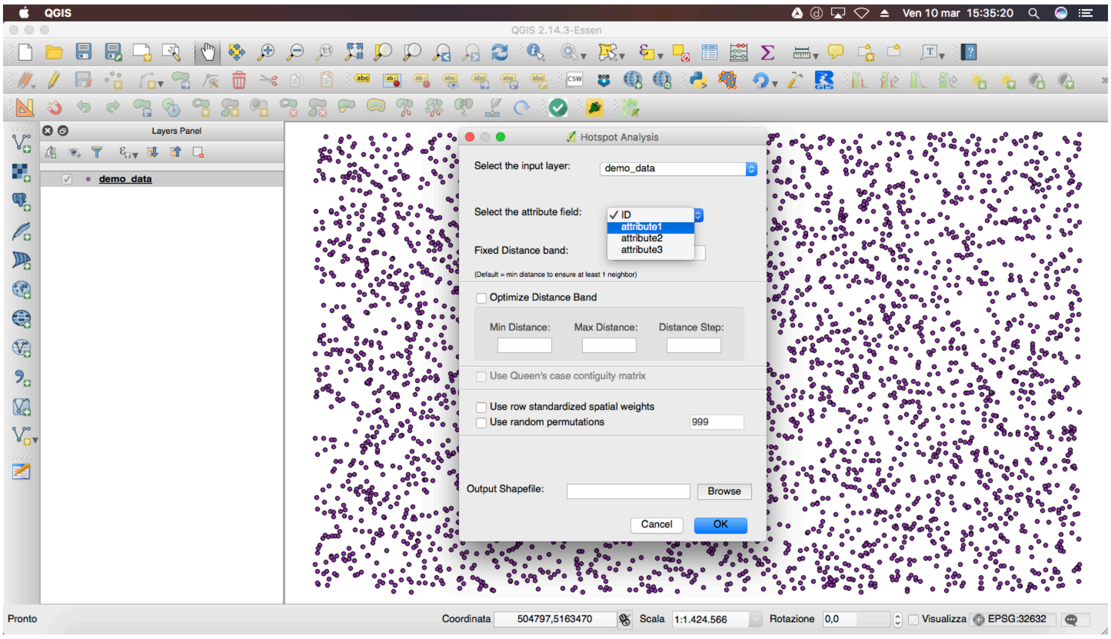
Sistema di Riferimento	ID dell'autorità
WGS 84 / UTM zone 29N	EPSG:32629
WGS 84 / UTM zone 29S	EPSG:32729
WGS 84 / UTM zone 2N	EPSG:32602
WGS 84 / UTM zone 2S	EPSG:32702
WGS 84 / UTM zone 30N	EPSG:32630
WGS 84 / UTM zone 30S	EPSG:32730
WGS 84 / UTM zone 31N	EPSG:32631
WGS 84 / UTM zone 31S	EPSG:32731
WGS 84 / UTM zone 32N	EPSG:32632

SR selezionato: WGS 84 / UTM zone 32N

```
+proj=utm +zone=32 +datum=WGS84 +units=m +no_defs
```

Attiva/disattiva modifica vettore in uso Coordinata 420706,5004932 Scala 1:1.413.278 Rotazione 0,0 ☐ Visualizza EPSG:32632

When the User Interface is open, select the input layer from the list as well as specify which attribute field contains the positive numerical attribute you want to use to run Hotspot Analysis. In the case of **demo_data** select one among: “*attribute*1”, “*attribute2*”, “*attribute3*”.



Hotspot Analysis

Select the input layer: demo_data

Select the attribute field: ✓ ID / attribute1 / attribute2 / attribute3

Fixed Distance band:

(Default = min distance to ensure at least 1 neighbor)

□ Optimize Distance Band

Min Distance: Max Distance: Distance Step:

□ Use Queen's case contiguity matrix

□ Use row standardized spatial weights

□ Use random permutations 999

Output Shapefile:

Pronto Coordinata 504797,5163470 Scala 1:1.424.566 Rotazione 0,0 ☐ Visualizza EPSG:32632

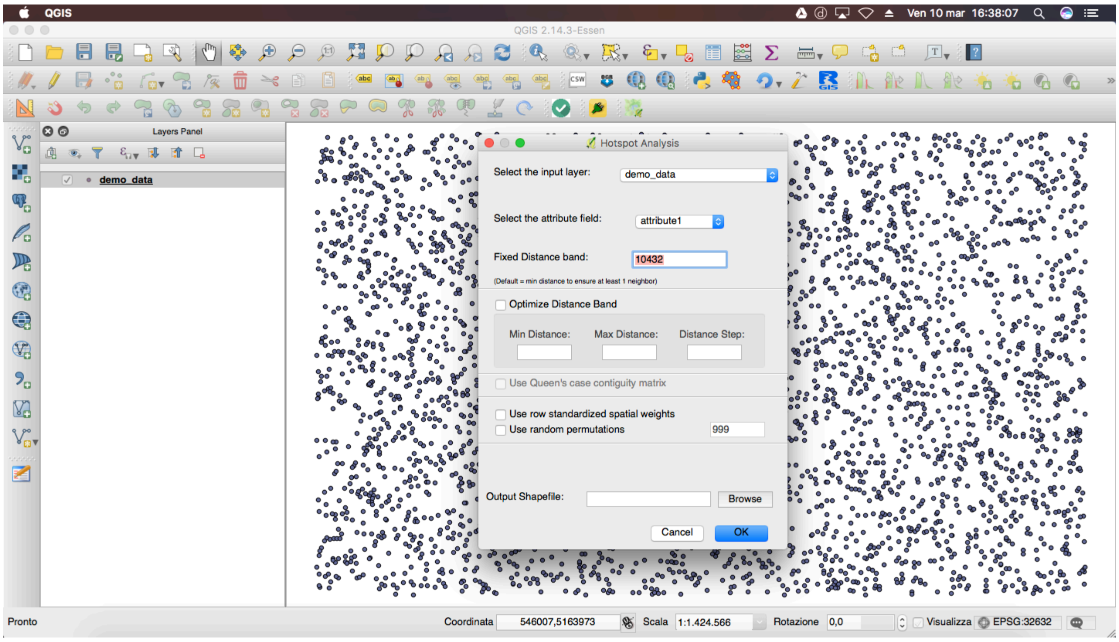
(Note: **demo_data** is an artificial dataset, attributes as well as point locations do not represent any physical feature)

1.2 Select a proper Distance Band for the analysis

As mentioned in the previous section, in order to compute local Getis-Ord Gi* statistic it is required to specify a spatial relation between points of the dataset. The plugin allows to modelled this relation using a **Fixed Distance Band**.
This process creates a PySAL **W** object (i.e. an adjacency or spatial weights matrix).
The Distance Band selection is a crucial step but no fixed rules are available to perform this task. Analysis distance depends directly on which phenomena your dataset describes as well as on the results you are looking for (e.g: if you are looking for hotspots of crime cases per city block within a city, your distance may be in the order of the average distance between two adjacent city blocks). Nevertheless, you should select a distance that guarantees some neighbours to any point (~ 8 for small datasets and ~ 30 for big datasets) in order to avoid so called “islands” and obtain reliable results.

The default suggested value for the Fixed Distance Band is the **minimum distance to guarantee at least one neighbour** to all the points of the dataset and it is derived by considering only their spatial distribution. You can change this value according to you needs but keep in mind that you should not select a lower distance.

In the case of **demo_data**, default suggested value for **Fixed Distance Band** is = 10423 m (according to the spatial distribution of points and the projected coordinate system of the layer, which is expressed in meters).



Hotspot Analysis

Select the input layer: demo_data

Select the attribute field: attribute1

Fixed Distance band: 10432

(Default = min distance to ensure at least 1 neighbor)

☐ Optimize Distance Band

Min Distance: Max Distance: Distance Step:

☐ Use Queen's case contiguity matrix

☐ Use row standardized spatial weights

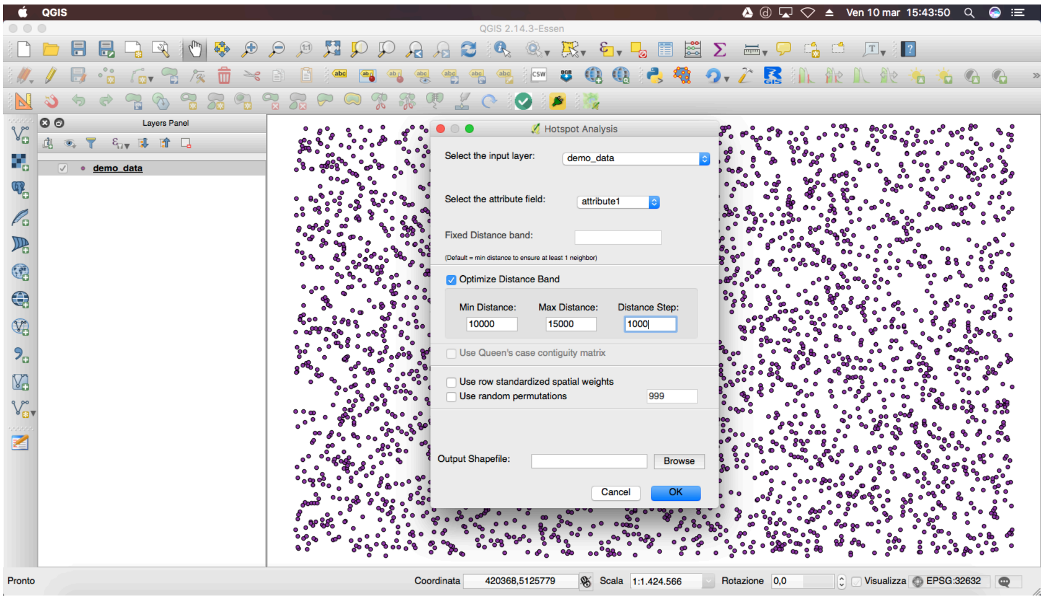
☐ Use random permutations 999

Output Shapefile: Browse

Cancel OK

The plugin includes also an optimization procedure in order to selected the distance band. To use this option, activate the command **Optimize Distance Band** by ticking the checkbox. Then you have to specify a distance range (by typing a minimum and maximum distance value) and a distance step

(example: **Min Distance** = 10000 m, **Max Distance**= 30000 m, **Distance Step**= 1000 m -> the algorithm will test this array of distance: [10000 m, 11000 m, …, 29000 m, 30000 m])



Hotspot Analysis

Select the input layer: demo_data

Select the attribute field: attribute1

Fixed Distance band:

(Default = min distance to ensure at least 1 neighbor)

☑ Optimize Distance Band

Min Distance: 10000 Max Distance: 15000 Distance Step: 1000

□ Use Queen's case contiguity matrix

□ Use row standardized spatial weights

□ Use random permutations 999

Output Shapefile:

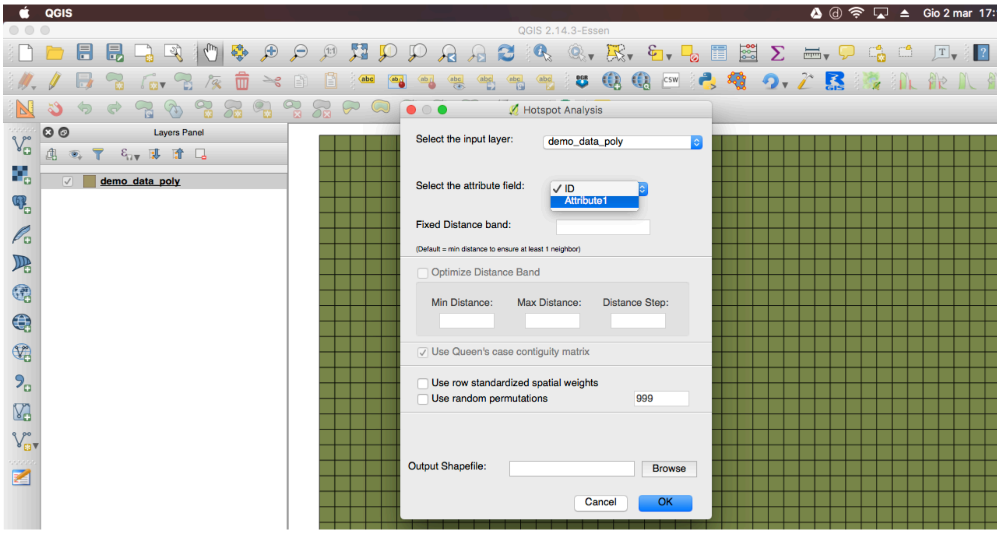
The optimized distance, which will be automatically used to perform Hotspot Analysis, is the one that maximize the Z-score of the **global Moran's I** index for your dataset.
This index suggests at which analysis distance the dataset shows high cluster activity (either of high or low values) and therefore is not only based on the points spatial distribution but also on the spatial arrangement of the associated variable (i.e. numeric attribute) that you select for the analysis.
Be aware that the suggested distance may not agree with your analysis purposes. You should specify a suitable range of distances to fit your analysis needs as well as you should not specify a Min Distance lower that the **minimum distance to guarantee at least one neighbour**. This is to avoid "islands" in the spatial weight matrix.

Computational time for the optimization increase according to the amount of points in your dataset but also is strongly related to the distance range and step you decide to adopt (i.e. **large distance range + little distance step = long computational time**!).

1.3 Polygon Shapefile

A demo input points shapefile is available here:
https://github.com/danioxoli/HotSpotAnalysis_Plugin/tree/master/test_data - > **demo_data_poly**
(Note: **demo_data_poly** is an artificial dataset, attributes as well as polygon locations do not represent any physical feature)

When your input layer is a polygons shapefile, the plugin will recognize it and it will deactivate the **Fixed Distance Band** options. This because spatial relationship between polygons is modelled using a first order **queen's case spatial weights matrix**. Therefore, two polygons are defined as neighbour when they share at least one vertex.
Also in this case, a valid input layer has at least a numeric attribute assigned to any geometry of the dataset. Important is also that geometries of your dataset are representative for a continue area avoiding holes (i.e. "island") in the spatial weights matrix.



Hotspot Analysis

Select the input layer: demo_data_poly

Select the attribute field: ✓ ID / Attribute1

Fixed Distance band:

(Default = min distance to ensure at least 1 neighbor)

□ Optimize Distance Band

Min Distance: Max Distance: Distance Step:

☑ Use Queen's case contiguity matrix

□ Use row standardized spatial weights

□ Use random permutations 999

Output Shapefile: Browse

Cancel OK

2. Plugin options and run

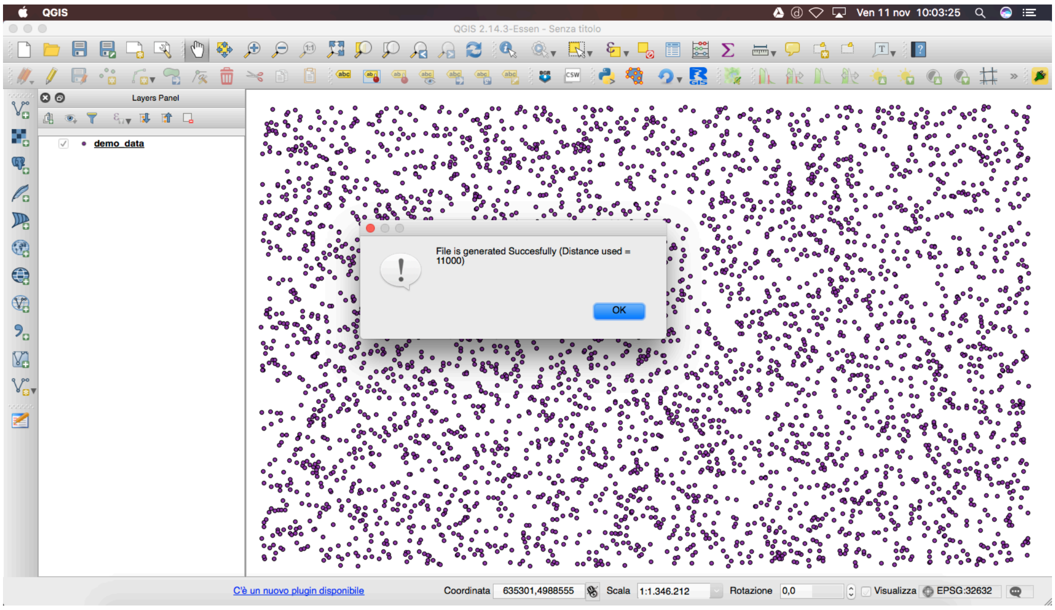
Once input layer, attribute and spatial relationship are defined you are ready to specify the output file name and path, you are ready to press **ok** and run the Hotspot Analysis!

Options are available to customize the GI* test according to your needs.
By activating the check-box “**Use row standardized spatial weights**” the spatial weights matrix is modified. Default weights are in fact binary, use row standardized weights means that any element of the matrix rows is divided by the sum of the correspondent row. This will balance in the computation the contributions locations characterized by high number of neighbours with location having only few neighbours:

$$W = \begin{Bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{Bmatrix} \xrightarrow[row\ standardization]{} \begin{Bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{Bmatrix}$$

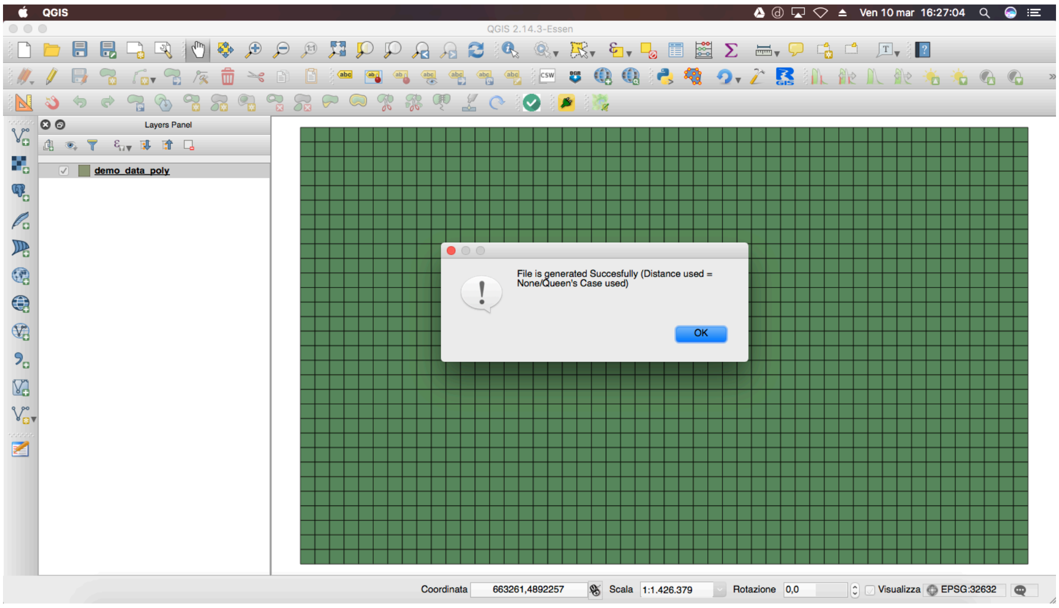
Another option available is the use of the permutation approach to compute the sampling distribution of the test statistic when the null hypothesis is true (null hypothesis = Complete Spatial Randomness). This can be activated through the check-box “**Use random permutations**”. Default number of permutation is 999, it can be changed by typing a number in the dedicated text box. When the random permutations are not invoked the default approximation for the sampling distribution of the test statistic is the Standard Normal Distribution. (In the example these two option are not activated)

At the end of the process a **success message** will be displayed and the output layer will be added automatically to the QGIS map panel. The success message includes also a reminder of the Distance Band you select. If the “Optimize Distance Band” option has been used, the distance reported is the optimized one that the algorithm selected for you. In case you are working with polygons layers, in the success message it will be specified that you have used a The generation of the output may take some time depending on the parameter you set as discussed before.



File is generated Succesfully (Distance used = 11000)

QGIS
QGIS 2.14.3-Essen - Senza titolo
Layers Panel
demo_data
OK
C'è un nuovo plugin disponibile
Coordinata 635301,4988555
Scala 1:1.346.212
Rotazione 0,0
Visualizza
EPSG:32632



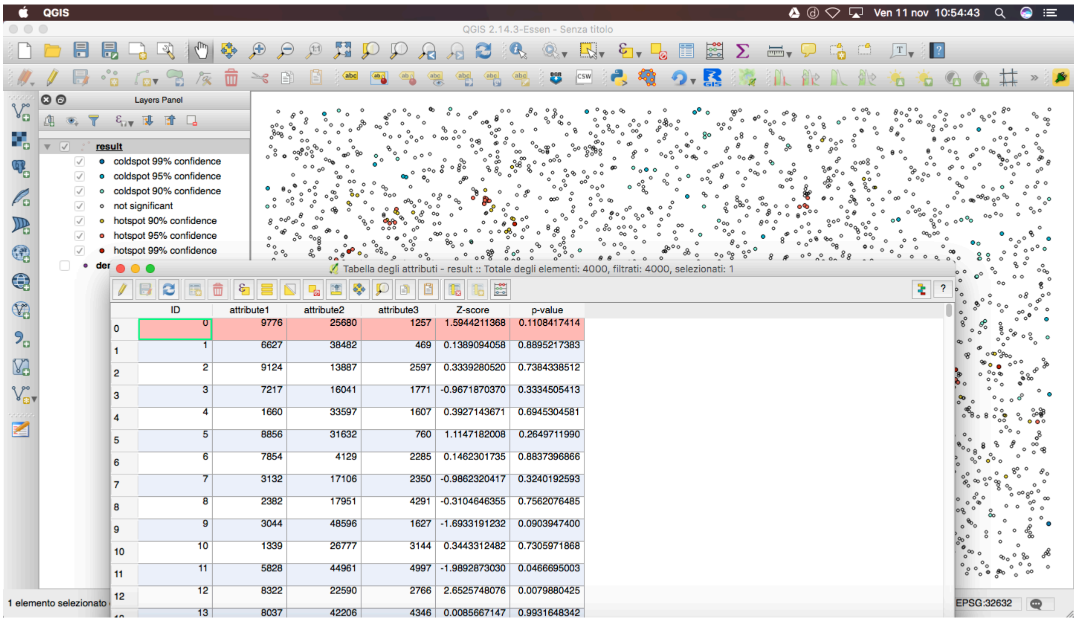
File is generated Succesfully (Distance used = None/Queen's Case used)

To the output layer is assign an automatic style which enable to distinguish between hotspots a coldspots, by accounting also for their statistical significance.

3. Interpretation of the results

Let's have a look to the attribute of the output layer.

As you it can be seen in the picture below, the output layer is a copy of the input layer with two new fields in the attribute table. These contains Z-scores of Gi* and related p-values for any entity of the dataset. The default style of the output layer uses combination of Z-score and p-values to distinguish between hotspots and coldspots while displaying their statistical significance. Threshold values are associated to the normal standard distribution as show in the following picture. Output styling can be changed according to user's needs.



QGIS 2.14.3-Essen - Senza titolo

Layers Panel

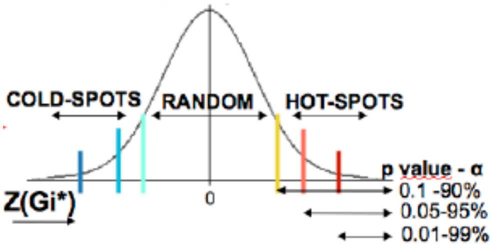
- result
  - ☑ ● coldspot 99% confidence
  - ☑ ● coldspot 95% confidence
  - ☑ ○ coldspot 90% confidence
  - ☑ ○ not significant
  - ☑ ● hotspot 90% confidence
  - ☑ ● hotspot 95% confidence
  - ☑ ● hotspot 99% confidence

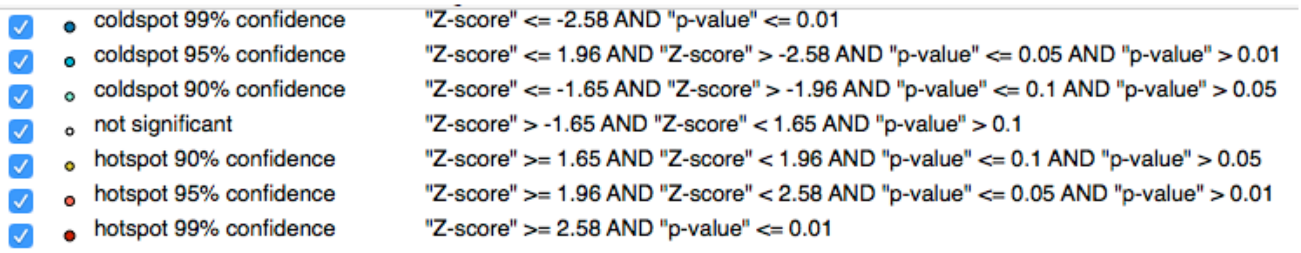
Tabella degli attributi - result :: Totale degli elementi: 4000, filtrati: 4000, selezionati: 1

	ID	attribute1	attribute2	attribute3	Z-score	p-value
0	0	9776	25680	1257	1.5944211368	0.1108417414
1	1	6627	38482	469	0.1389094058	0.8895217383
2	2	9124	13887	2597	0.3339280520	0.7384338512
3	3	7217	16041	1771	-0.9671870370	0.3334505413
4	4	1660	33597	1607	0.3927143671	0.6945304581
5	5	8856	31632	760	1.1147182008	0.2649711990
6	6	7854	4129	2285	0.1462301735	0.8837396866
7	7	3132	17106	2350	-0.9862320417	0.3240192593
8	8	2382	17951	4291	-0.3104646355	0.7562076485
9	9	3044	48596	1627	-1.6933191232	0.0903947400
10	10	1339	26777	3144	0.3443312482	0.7305971868
11	11	5828	44961	4997	-1.9892873030	0.0466695003
12	12	8322	22590	2766	2.6525748076	0.0079880425
13	13	8037	42206	4346	0.0085667147	0.9931648342

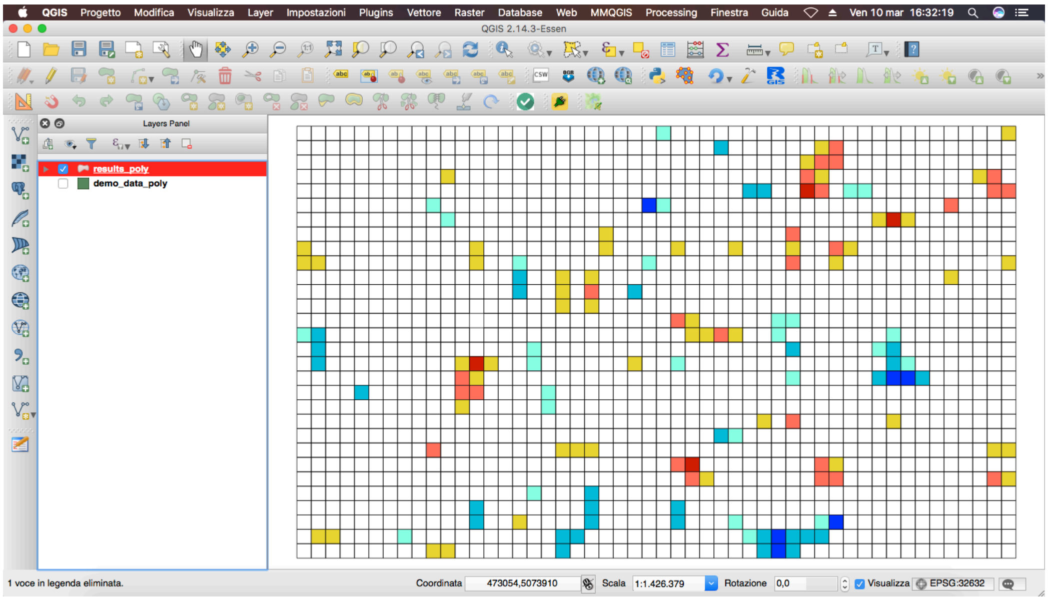
1 elemento selezionato

EPSG:32632





☑	●	coldspot 99% confidence	"Z-score" <= -2.58 AND "p-value" <= 0.01
☑	●	coldspot 95% confidence	"Z-score" <= 1.96 AND "Z-score" > -2.58 AND "p-value" <= 0.05 AND "p-value" > 0.01
☑	○	coldspot 90% confidence	"Z-score" <= -1.65 AND "Z-score" > -1.96 AND "p-value" <= 0.1 AND "p-value" > 0.05
☑	○	not significant	"Z-score" > -1.65 AND "Z-score" < 1.65 AND "p-value" > 0.1
☑	●	hotspot 90% confidence	"Z-score" >= 1.65 AND "Z-score" < 1.96 AND "p-value" <= 0.1 AND "p-value" > 0.05
☑	●	hotspot 95% confidence	"Z-score" >= 1.96 AND "Z-score" < 2.58 AND "p-value" <= 0.05 AND "p-value" > 0.01
☑	●	hotspot 99% confidence	"Z-score" >= 2.58 AND "p-value" <= 0.01



QGIS 2.14.3-Essen

Layers Panel

- ☑ results_poly
- □ demo_data_poly

1 voce in legenda eliminata.

Coordinata 473054,5073910 Scala 1:1.426.379 Rotazione 0,0 ☑ Visualizza EPSG:32632

Hotspots represent atypical high-value location surrounded by other high-value location as well (and coldspots vice-versa). Not significant points represent location in which local values are likely random distributed and so no significant clusters are there located.
This process enables to describe and visualize spatial distributions by highlighting atypical locations which can be fundamental to describe hidden patterns of your dataset.

Daniele Oxoli
Ph. D Student, Politecnico di Milano
Polo Territoriale di Como, **GEOLab**
Email: daniele.oxoli@polimi.it
11/09/2016