Spatial Data – the Final Frontier

April 2009

This month's topic: A Generalised Repository

The Department has masses of spatial data, of differing scales and quality, which has been captured at different times for different purposes. We also frequently have the need for background data for orientation purposes – to be placed behind specific data to locate and orient that data.

There are significant costs incurred whenever spatial data is "borrowed" for this kind of use, and this cost is added to the cost of the research. It is possible, with very little outlay, to build a repository which will largely eliminate these costs.

This concept is based on work that has been carried out by:

> ITC, International Institute for Geo-Information Science and Earth Observation, Enschede, The Netherlands.
> The Office of the Deputy Prime Minister, UK.
> The Ordinance Survey, UK.
> Department of Geography, University of Zürich.

Note that the sort of process being suggested here is already in operation, and producing these benefits in the UK.

It might be thought that duplication of features, as suggested here, would be space consuming, and lead to update duplication, but this is not the case. The storage requirements are very modest (about 33% increase over a non-generalised repository), and there is no requirement for manual intervention in the feature generalisation process.

Detail

The department's databases contain a collection of data at varying scale, and of varying quality. Also, in most cases, metadata is available to determine the usefulness of that data for various purposes.

On the other hand, the database it is not complete. There is complete coverage of most themes only at the smaller scales – such as 1:250k[1]. The larger scale data is only available over certain areas and for some feature types. This is not a criticism, and there is no possibility of ever "completing" a full set of coverages, but there is an opportunity to extend the existing cover to make a more complete product.

For example, there are areas with no 1:50k data, where there is coverage at 1:25k. The technology now exists to produce reasonable scale-reduced data, so that the 1:50k coverage can be extended. (Naturally, the reverse is not possible).

---

[1] A scale of one to 250,000. That is: 1 cm on the page is equivalent to 250,000cm (or 2.5km) on the ground. The term "small scale" is used for cases where a small amount of paper (or screen space) is required to represent a region. Large scale requires a larger sheet of paper. Thus 1:250,000 is small scale, 1:2500 is large scale.

Functionality is available which will provide background mapping data at ALL scales from (say) 1:2m to 1:2,500, with complete coverage at the small end, and as much coverage as possible at the large scale end of the range.

Classification	Scale 1:	From 1:	to 1:	contour interval (m)	contour index interval (m)	resolution (m)
E	5m	9,999,999	5,000,000	200.000	500.000	1,000
F	2m	4,999,999	2,000,000	200.000	500.000	500
G	1m	1,999,999	1,000,000	150.000	450.000	250
H	500k	999,999	500,000	100.000	500.000	100
I	250k	499,999	200,000	50.000	250.000	50
J	100k	199,999	100,000	20.000	100.000	25
K	50k	99,999	50,000	10.000	50.000	10
L	25k	49,999	20,000	5.000	25.000	5
M	10k	19,999	10,000	5.000	10.000	2.5
N	5k	9,999	5,000	2.500	5.000	1.0
O	2,500	4,999	2,000	1.000	5.000	.5

Table 1 – Scale Classifications used in RIME

As an example, RIME uses the concept of scale classification, which is, in effect, an indication of the scale, and has a classification for approximately every doubling of scale. This provides for all the standard scales, and also for special purpose scales (and the old imperial scales). The letters A to D and P to Z are reserved for expansion.

Requirements

In broad terms, the requirement is for data to be made available at all scale classifications, with fast response, and in the most useful format(s).

There are two issues to be addressed where data of larger scale is to be used in a smaller scale form. Together these techniques are referred to as “generalisation”:

- The most appropriate features should be selected, and the less appropriate ones dropped.
- The amount of detail should be reduced in the features that are retained.

Software for the generalisation process exists from various sources, but is constantly being developed and improved. The question of detailed requirements for this software has been accepted as an evolving field of research. So that this part of the requirements statement must remain “as good as we can get”.

On the other hand, very acceptable results are being obtained in Europe.

Issues of what types of features are of interest, and whether the data should be 2D or 3D will need to be refined as the process proceeds, and will be likely to change as it evolves.

Repository

In effect, the data would be stored in a repository, with a table of features for each of the appropriate scale classifications. For example, a feature captured with an equivalent scale of 1:25k is determined to have a maximum scale classification of L.

While spatial data may not have an actual scale, all spatial data has accuracy characteristics that determine the largest scale at which it is meaningful. Provided this is known, the data can be included in the generalised repository.

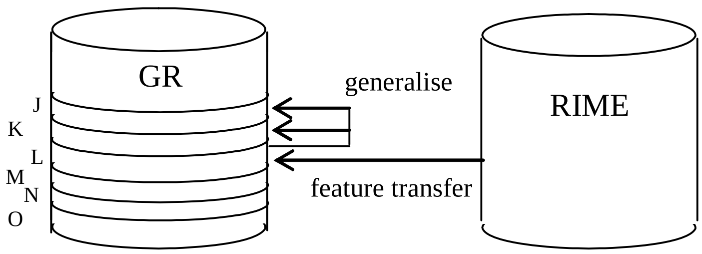


Figure 1 One option for the process of creating a Generalised Repository

The feature could be copied into the L (1:25k) layer, generalised to 1:50k and a copy placed in the K layer. A further generalisation to 1:100k can be placed in the J layer, and the process continued until better data of that theme is available, or the limit of generalisation is reached.

Use might be made of the experimental web generalisations services provided by the University of Zurich.

Presentation

The generalised repository would make available a number of services (e.g. Web Map Service - WMS, Web Feature Service - WFS), and also serve up GML (Geographic markup language) and shape files etc. The presentation of the data from those services would be independent of this process. Clients of various kinds could be accommodated – from web map clients to GIS's to Google Earth.

The levels of generalisation mean that the amount of work needed by a server at any map scale is independent of scale. Thus a simple enquiry such as a web map server enquiry will operate at a speed that does not depend on the zoom level..

The time to service a request will depend only on the actual amount of data needed. For example, the data needed to build a "smart map' sized keymap will be sub-second, but the time to serve up data to produce a 1 metre square map at appropriate scale may take a minute or two, but the speed will not depend strongly on the area of interest.

Storage and Update

The storage requirements of the method are quite modest. If an area is covered by features of a specific scale, but not the next smaller scale (e.g. if 1:100k data is available, but not 1:250k), the number of features required for the smaller scale data will be about ¼ of the original number.

This is only approximate, but the effect is enhanced by the generalisation routines, which means that each feature will need to be less detailed – theoretically only ½ the number of points are required. In practice, because neither the weeding nor the generalisation routines are perfect, it can be expected that a set of features, weeded and generalised by one classification will require about ¼ of the original space.

If the process continues, the space will be $1+\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+...$ which sums to about 1.33%.

Progress So Far

A small area of RIME data has been split into the scale classifications as described above. The chosen area is over the northern suburbs of Brisbane, and as a result, data at all scales between 1:5m and 1:2,500 is available.

Very primitive methods of selection of features have been used in the generalisation process, and this is obvious from the results.

No attempt has been made to reduce the detail of individual features. (Thus the service times are slower than can be expected in the final system.

A basic web map server has been prepared, so that the data can be viewed.

This has proven the concept, with very acceptable WMS speeds over this limited area, and has shown that sub-second response times are readily achievable at every scale level in a full sized database.

(See attached image samples).

Preliminary Results

On the subsequent pages, a series of images have been generated from the "proof of concept" generalised repository. The images include about the level of detail that would be expected of a web map service, but it should not be inferred that this is the aim of the proposal. A web map service is only one minor part of the benefits to be obtained by this approach.

The images are presented in pairs, with the right image being a "zoomed in" version of the left. The left image on the next pair is (usually) the same area and scale as the prior image, but using data from the next level of detail. Thus it can be imagined that a continuous zoom is being done, over the complete range of sensible map scales.

The images are to be interpreted as though they were screen "keymaps", not as publishable maps. They were generated using web map server software. There are several obvious flaws in the images (such as missing sections of the Brisbane River), which are a result of the primitive feature-weeding system that was used in the demo. An improvement in this, using available techniques, and the introduction of a generalisation algorithm would result in an improved product.

In the early stages, this approach would bring us the functionality of web-map services, Google Earth type operations, and topographic data in Smart Map, but it can be carried to much higher quality levels – with publication-grade data at any requested map scale (as is available in the UK).

Conclusions

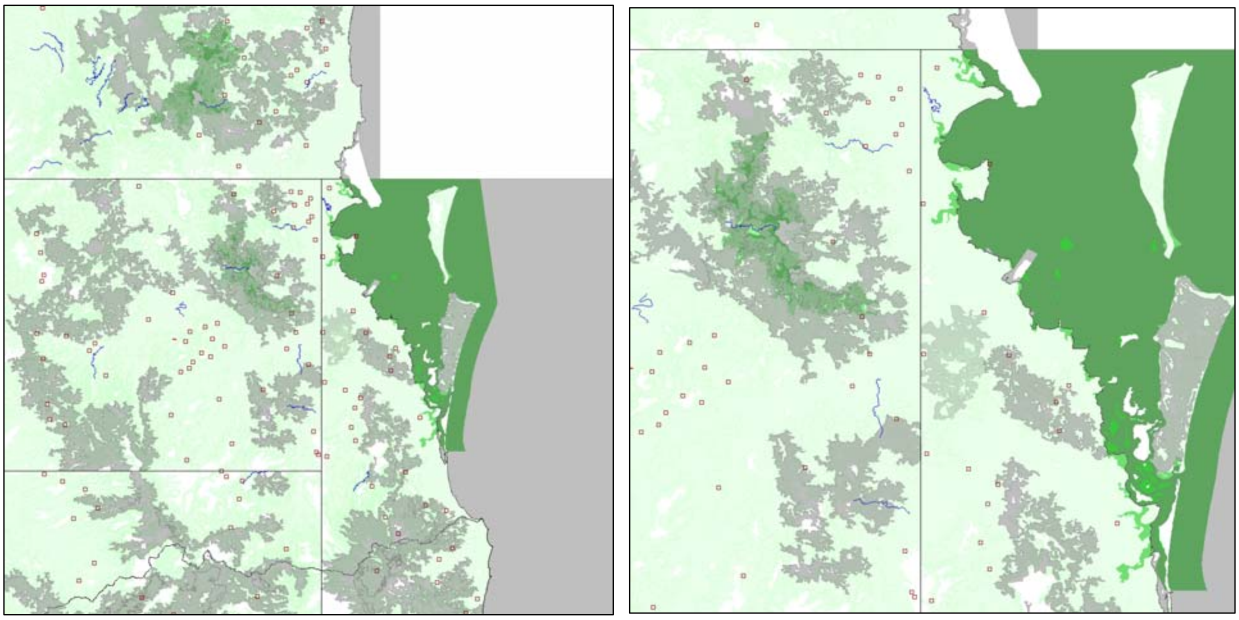
The technology is available, and the British Ordnance Survey has done something of this sort with their “Master Map”.

This cannot be achieved by a once-only “big bang” project. It needs cooperative input from data custodians, IT and cartographers to be able to work. The details of the feature-weeding and generalisation rules will evolve as the data is enhanced, and change as new techniques become available.

Rod Thompson

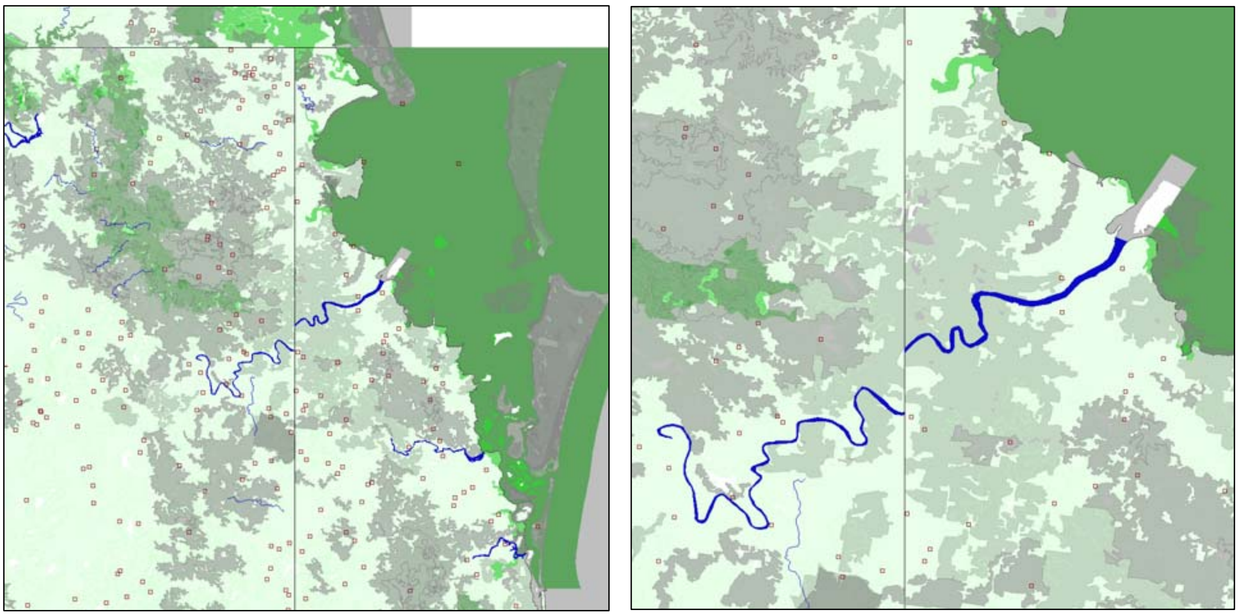
April 2009

From Scale Classification E,effective scale 1:5m



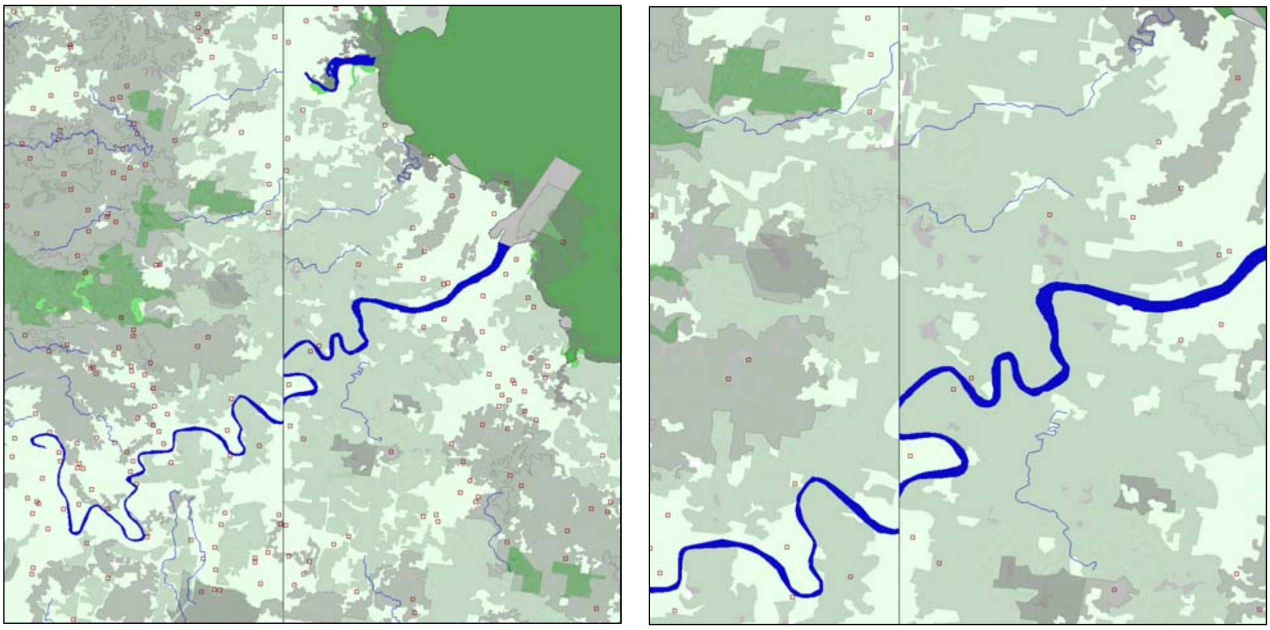
This level is generalised from 1:250k data, by removing smaller area features, shorter linear features, and a random selection of point features. Note that rivers have broken up into random segments. This would not have happened if they had been named or strung together as single features.

From Scale Classification F, effective scale 1:2.5m



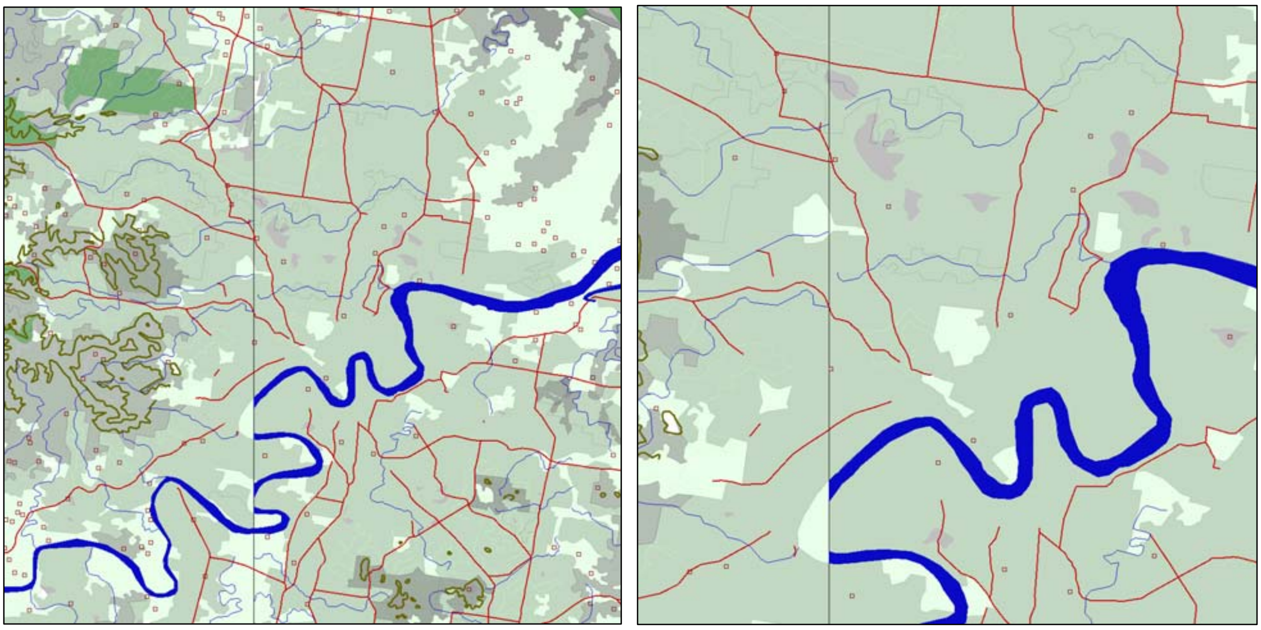
This level is also generalised from 1:250k data, by removing smaller area features, shorter linear features, and a random selection of point features. Note the break in the Brisbane River – This is an artefact of the area feature being broken at a sheet boundary in the data. Again, this will be corrected automatically when the feature is named.

From Scale Classification G, effective scale 1:1m



This level is also generalised from 1:250k data, by removing smaller area features, shorter linear features, and a random selection of point features. Note the break in the Brisbane River. The break can be seen here to be caused by missing small segments rather than an edge mis-join.

From Scale Classification H, effective scale 1:500k

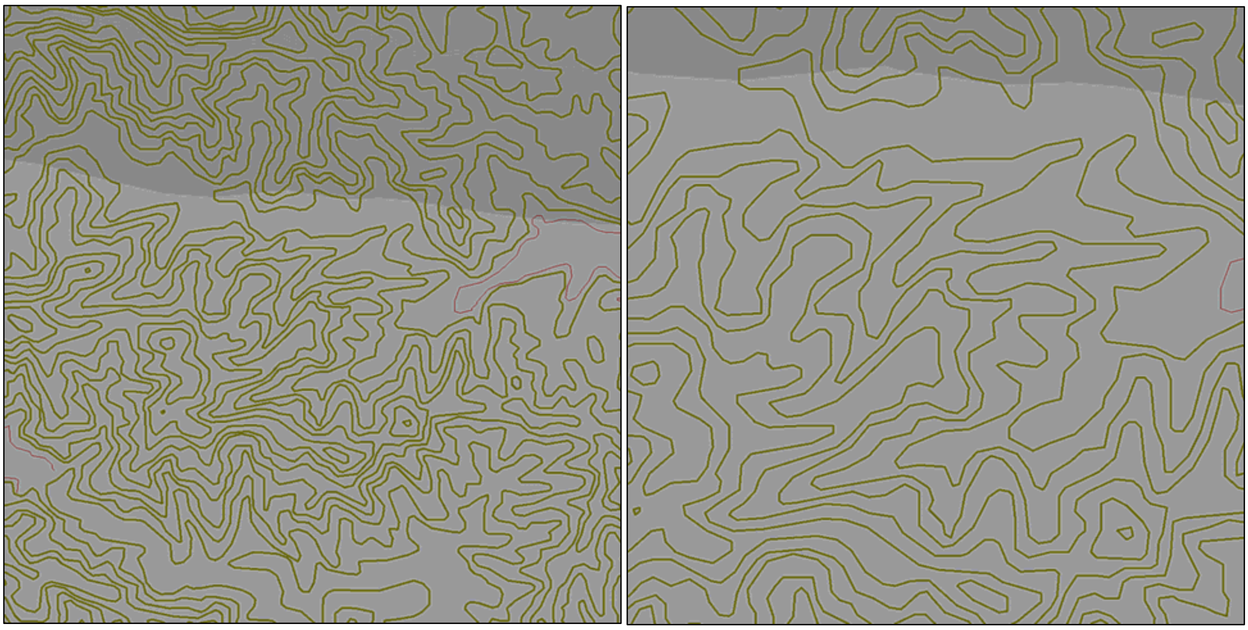


This level is also generalised from 1:250k data, by removing smaller area features, shorter linear features, and a random selection of point features. Note the break in the Brisbane River – the break can be seen here to be caused by missing small segments rather than an edge mis-join.

From Scale Classification I, effective scale 1:250k

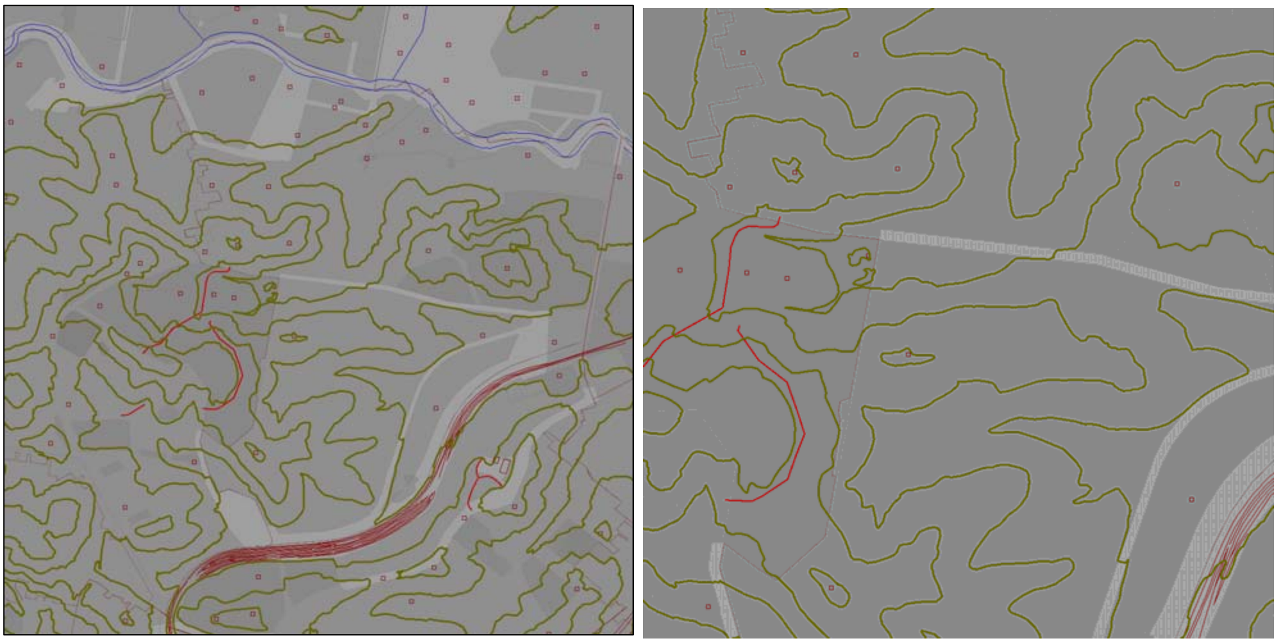
This is unedited 1:250k data. It was also used as the source for the levels E to H. Note the bight of the river that was removed form the smaller scale data – in particular, compare with level H.

From Scale Classification J, effective scale 1:100k



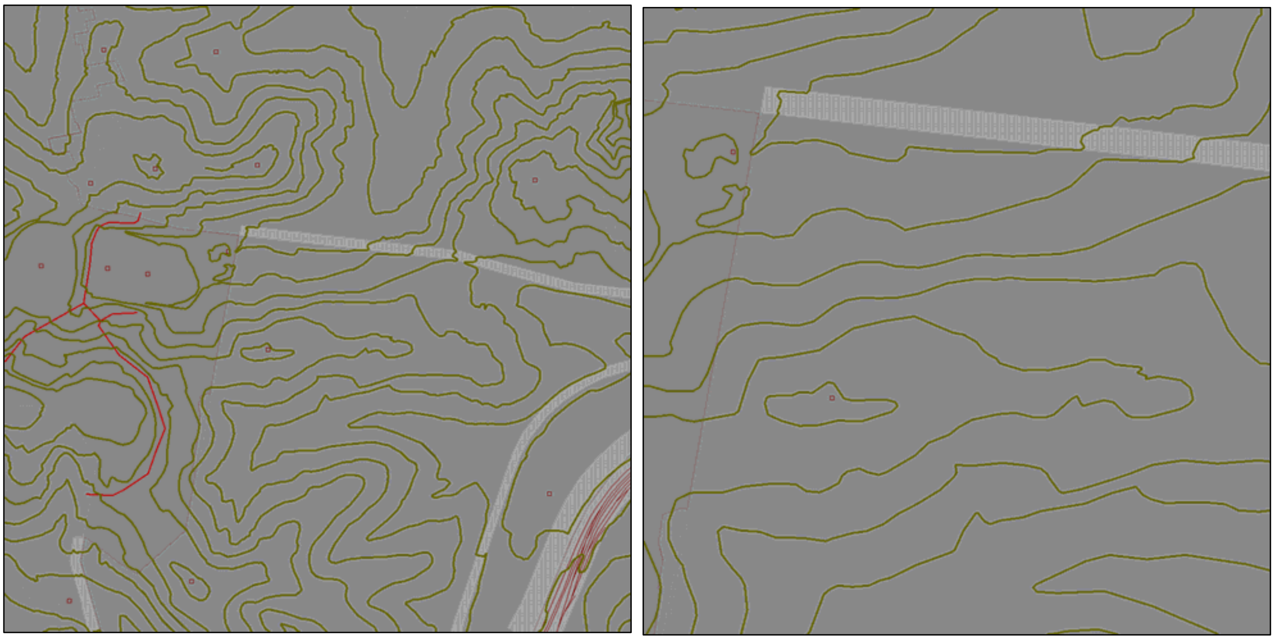
This is source 1:100k data. Note that this is a different centre point to the previous level (I). At the area where the previous regions were centred there is no large scale data available.

From Scale Classification K, effective scale 1:50k



This has been taken from 1:25k data, omitting all contours and index contours that are not divisible by 10m.

From Scale Classification L, effective scale 1:25k



This is original 1:25k data.

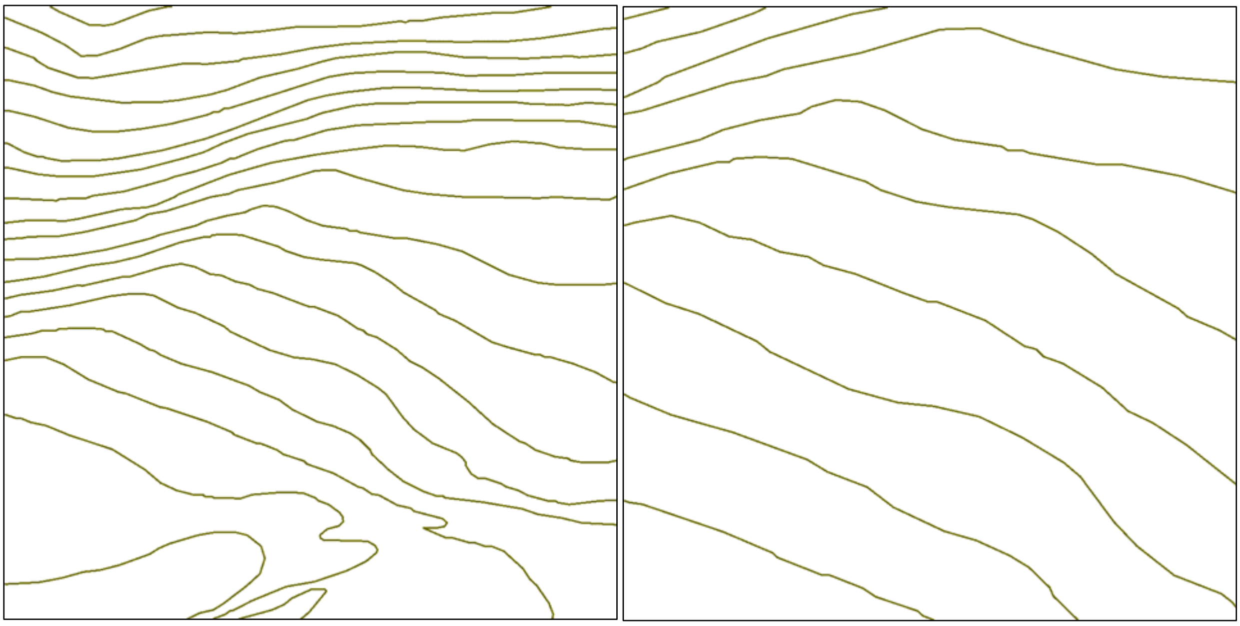
From Scale Classification M, effective scale 1:10k

This is taken from 1:2500 data (classification O), omitting all contours (leaving index contours).

From Scale Classification N, effective scale 1:5k

This is taken from 1:2500 data (classification O), omitting all contours and index contours with an elevation not divisible by 2m.

From Scale Classification O, effective scale 1:2500



This is the original 1:2500 data.