Stockholms Universitet
Statistiska institutionen
Daniel Thorburn

Tentamen i statistiska databaser och register (Statistical Databases and registers)
.

20 mars 2009, kl. 9.00-14.00 i Ugglevikssalen.

The test contains of two parts 1) Infological aspects on databases and 2) Data mining and other matters. **The answers and solutions of the two parts should be written on separate papers.** The examination consists of these two parts and the oral exam on the part Statistical aspects and quality of registers. You must pass all three parts to pass the course. The final mark is based on a weighting of the thre parts, Infology, Data mining and Register quality in the proportions 2:1:2.

All assumptions and notations should be explained and defined (also those that have been used during the course). All answers, reasoning and explanations should be easy to follow. Answers and arguments which cannot be understood give 0 p.

The tests will be handed back March 30 at 10 am in B 705 .

First part: Infological Aspects on Databases

Look at the graph below, and answer the following questions:

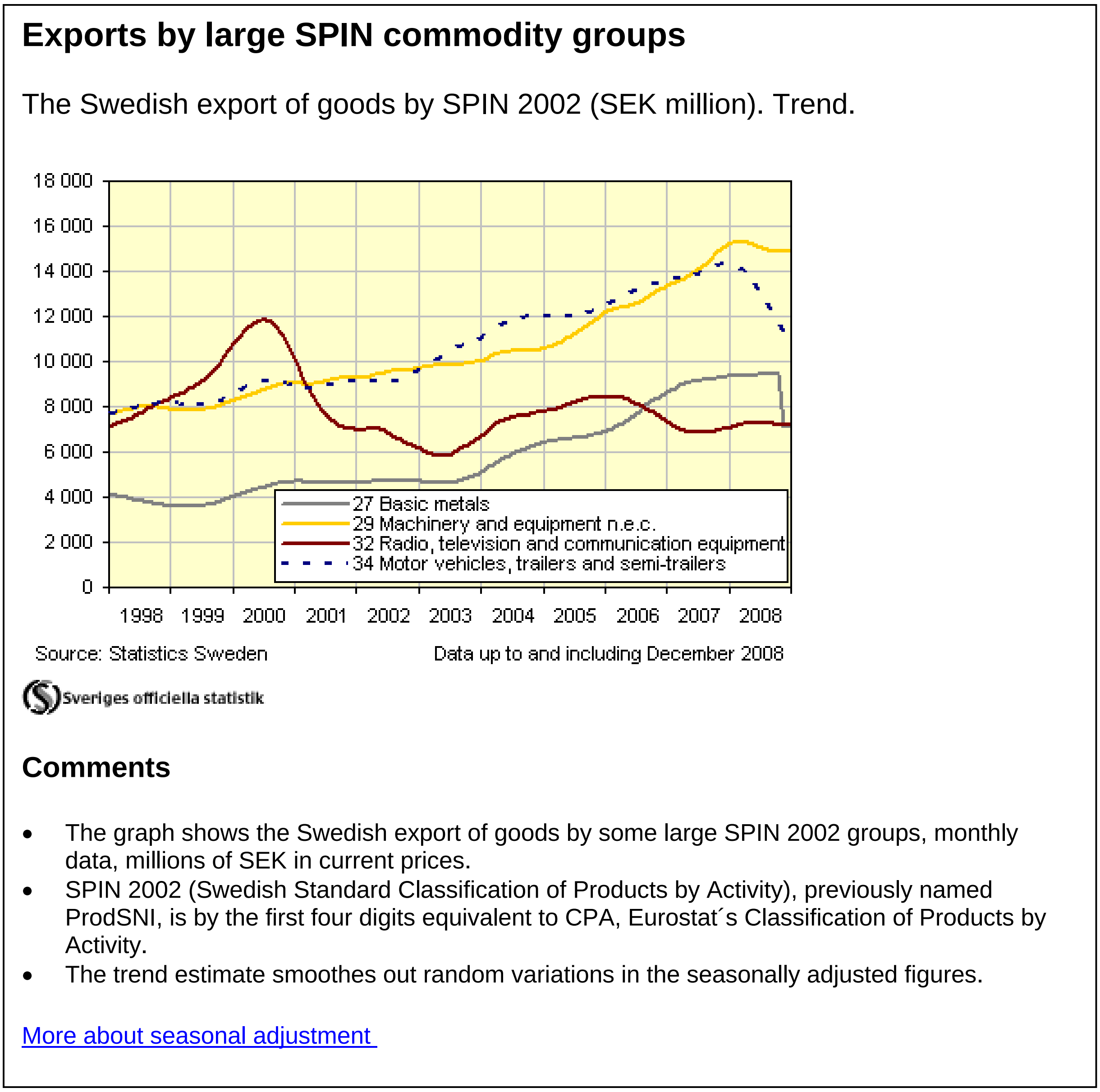
1. The graph (and an equivalent statistical table) may be assumed to be based upon an aggregation of trade transaction data. Try to reconstruct a conceptual model (data model) of the underlying transaction data, indicating relevant object types, variables of the objects, and relations between the object types. Use the modelling technique and symbols that you saw many examples of during the course, and which you used in an exercise. Note that there are several solutions to this problem that may be correct under different assumptions. You are free to choose any of these solutions, but please write down any assumption that you make

2. In the example below (the graph and the surrounding texts), you can see both data and metadata. Explain and exemplify what is data and what is metadata in this example?

3. Metadata are needed for different purposes in connection with statistical databases and statistical information systems. Mention at least three different purposes for which metadata are needed. Mention also at least three different categories of users of metadata.

4. Give some examples of metadata, which cannot be found in the example above, but which may be necessary to have for certain purposes. For which purposes and users may these ”missing” metadata be needed, and where could they be available?

5. The statistics in the graph below could be stored in a so-called multidimensional cube (hypercube). Describe the structure of such a cube for this example, and indicate in particular which the dimensions of the hypercube would be, and what would be the contents of the cells in the hypercube.



Exports by large SPIN commodity groups

The Swedish export of goods by SPIN 2002 (SEK million). Trend.

18 000
16 000
14 000
12 000
10 000
8 000
6 000
4 000
2 000
0
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
27 Basic metals
29 Machinery and equipment n.e.c.
32 Radio, television and communication equipment
34 Motor vehicles, trailers and semi-trailers
Sveriges officiella statistik

Source: Statistics Sweden

Data up to and including December 2008

Comments

- The graph shows the Swedish export of goods by some large SPIN 2002 groups, monthly data, millions of SEK in current prices.
- SPIN 2002 (Swedish Standard Classification of Products by Activity), previously named ProdSNI, is by the first four digits equivalent to CPA, Eurostat´s Classification of Products by Activity.
- The trend estimate smoothes out random variations in the seasonally adjusted figures.

More about seasonal adjustment

Second part: Data mining and other matters.

1. During the course two types of analysis were specially mentioned – Pattern recognition and Predictive analysis. Explain the terms and exemplify.

2. Explain the terms training data set, test data set and scoring data set and describe how these concepts are used in datamining and statistical analysis.

3. Describe and exemplify the use of decision trees and logistic regression when trying to find valuable rules for predicting a binary variable.

4. What is a neural network and when is it used?

5. Give a short account of the big private actors working on the business register market databases. Who are they and what do they do?