ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY
Lönnrotinkatu 4 B, 00120 Helsinki 12, Finland, tel. 601322

Keskusteluaiheita
Discussion papers

Jukka Lassila

PRELIMINARY DATA IN ECONOMIC DATABASES

No 306 10.11.1989

To be presented at the 5th International Conference on Statistical and Scientific Databases, Charlotte, North Carolina, April 3-5, 1990.

Forthcoming in Lecture Notes in Computer Science by Springer-Verlag.

ISSN 0781-6847

LASSILA, Jukka, PRELIMINARY DATA IN ECONOMIC DATABASES, Helsinki : ETLA, Elinkeinoelämän Tutkimuslaitos, The Research Institute of the Finnish Economy, 1989. 10 p. (Keskusteluaiheita, Discussion Papers, ISSN 0781-6847 ; no. 306).

ABSTRACT: Old preliminary data should not be removed when new data arrives, otherwise important information is lost. Economic research, especially modelling, needs this data. The costs of saving old preliminary data can be made very small by proper arrangement of data and with simple bookkeeping about when and from where the data came. As an example, the database of the Research Institute of the Finnish Economy (ETLA) is described.

KEY WORDS: preliminary data, source information

LASSILA, Jukka, PRELIMINARY DATA IN ECONOMIC DATABASES, Helsinki : ETLA, Elinkeinoelämän Tutkimuslaitos, The Research Institute of the Finnish Economy, 1989. 10 s. (Keskusteluaiheita, Discussion Papers, ISSN 0781-6847 ; no. 306).

TIIVISTELMÄ: Vanhoja ennakkotietoja ei pitäisi poistaa tietokannasta uudempien tietojen tullessa. Muuten menetetään tärkeätä informaatiota, jota voitaisiin käyttää tutkimuksessa ja mallien rakentamisessa. Vanhojen ennakkotietojen säilytyksen kustannukset saadaan hyvin pieniksi järjestämällä aineisto sopivasti. Lisäksi täytyy pitää kirjaa siitä, milloin ja mistä lähteestä aineisto on tullut. Esimerkkinä tarkastellaan Etlan tietokantaa.

ASIASANAT: ennakkotiedot, lähdeinformaatio

PRELIMINARY DATA IN ECONOMIC DATABASES

CONTENTS

Preliminary data in economic databases

Abstract: Old preliminary data should not be removed when new data arrives, otherwise important information is lost. Economic research, especially modelling, needs this data. The costs of saving old preliminary data can be made very small by proper arrangement of data and with simple bookkeeping about when and from where the data came. As an example, the database of the Research Institute of the Finnish Economy (ETLA) is described.

1. Introduction

Most of the newest economic figures are preliminary. Almost all macroeconomic data is. It is also by far the most interesting data if one is monitoring the economic situation. Producers of commercial economic database services around the world hurry to put it into their computers.

Things begin to go wrong when new revised data appears. The quality of the data is probably better, although revised estimates are still in many cases preliminary and will change in the future. The problem is that in most economic databases old preliminary figures are deleted when new revised data is saved.

The principle, thus, is for each variable to have only one observation for each point in time or period. This is a clear principle, and clearness is its main merit. Confusion would be a much worse alternative. Still, revising Einstein's famous remark, "things should be made as clear as possible, but not more so." Changes and revisions are the hallmark of economic data, and databases should provide information about the nature of these changes.

There are studies concerning the properties of preliminary statistical figures and their use in e.g. model building, see (1), (2) and (3). And there is a whole body of literature concerning a technically very closely related issue, forecast errors, see (4). Still, information about preliminary data is scarce compared with the volume of that data.

The solution proposed here is simple. Old preliminary data should not be removed when new data is entered into databanks. Users of new data may then browse older data and thus get a feeling about the frequency and magnitude of revisions. Statistical studies concerning preliminary data would be much easier and cheaper to

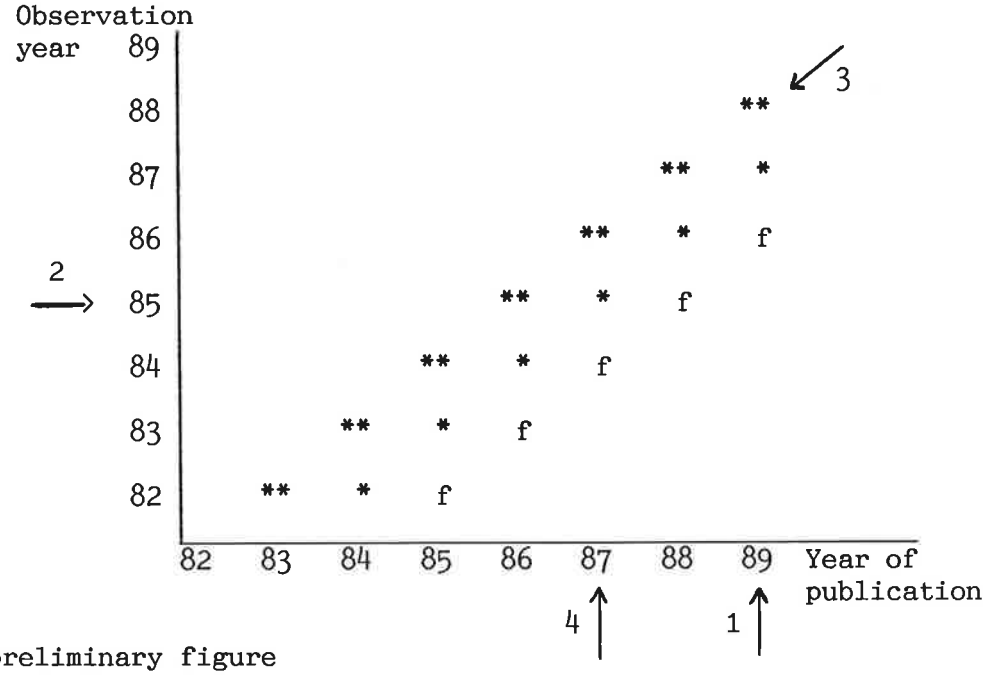
carry out. The proposed solution can be arranged easily, as almost a by-product of normal data updating procedures.

Section 2 presents a slightly simplified data model, and the rest of the sections describe the structure of ETLA's (the Research Institute of the Finnish Economy) database.

2. A model of preliminary data structure

Let us assume that statistics on production are published once a year, and that each publication contains figures for three preceding years. The figures are either preliminary, revised or final, according to the following pattern.

Figure 1.



** = preliminary figure
* = revised preliminary figure
f = final figure

There are four types of questions that are typically asked from the data in Figure 1. They are related to different ways of looking at the data, and the ways are indicated by the arrows.

Question 1. What is known about production, now?

This is the question that the commercial databases are best suited to answer. The user is e.g. a business economist who wants the most up-to-date figures concerning production. The answer consists of final figures for 82 - 86, a revised preliminary figure for 87 and a preliminary figure for 88.

Question 2: How did preliminary figures concerning production in 85 converge to the final figure?

This may be a research question, a case of reliability of preliminary statistical figures concerning a specific point in time. Or, a business economist, aware of the fact that revisions occur, wishes to have a feeling for previous changes to be able to better judge the uncertainty in newer preliminary figures.

Question 3: How good have first preliminary figures been in relation to final figures?

This may also be a research question, but now in a time-series sense. The aim may be to study the (un)biasedness of preliminary figures. Or again, a business economist is trying to grasp the magnitude of data revisions, while his main interest is focused on the newest figure.

Most economic databases cannot answer questions 2 and 3, and have never been able to answer them.

Notice that arrow 1 is pointing toward the year 89 (now). It could as well be pointing toward 87, corresponding to a slightly different question:

Question 4. What was known in 87 about production?

The reasons for asking question 4 are probably vastly different from those for asking question 1. Question 4 may arise in connection with research concerning the making of economic policy, e.g. why was monetary policy eased in autumn 87?

Most databases cannot answer question 4 anymore, even though they probably could have answered it in 87, when 87 was 'now'.

The amount of information necessary to answer questions 2,3 and 4 is not large: all the preliminary, revised and final figures and the time when each observation was published. When data structures are more complex than in the example above, one usually needs to know also where each observation came from.

3. Preliminary figures in ETLA's database

The main differences between reality and the above simplified data model are:

- There are many producers of figures. This is usually not the case with historical statistics, but it is very typical of forecasts. There may be hundreds of institutions predicting U.S. inflation.

- The publication pattern is not regular in practice. In some years more publications come out or figures may even be revised for the whole history at the same time.

- Definitions and measuring rules change from time to time. Base years change. New series emerge and old ones are dropped. In this sense, data structures do not remain constant for long periods.

ETLA's database mimics the statistical publications from which the data is updated. Each table in the database is updated from only one statistical source. Thus in any table the source of all the figures in any one row is the same. This is an essential feature of a system with old preliminary data available alongside new figures.

If database tables were combined from several statistical sources, it would be much more cumbersome to code when and from where each observation (each row of figures) came.

This has turned out to be a working solution also from an updating point of view: the whole table is updated at the same time. Documents to help updating are easily produced: a copy of the statistic is a document of the database table.

One table is updated from only one source, but one source may provide data for many tables, of course. National Accounts Statistics, e.g. is divided into about one hundred tables.

A database table is a matrix. A typical table is as follows, with QQ, CQ and IQ referring to total production, total consumption and total investment respectively.

Table 1.

"OBS"	"SID"	"STATUS"	"QQ"	"CQ"	"IQ"
"Y:1983"	"NA8710"	"X"	315249	234443	79468
"Y:1984"	"NA8710"	"X"	325505	240910	77786
"Y:1985"	"NA8710"	"X"	336824	249882	80052
"Y:1986"	"NA8710"	"-"	344996	259396	79110
"Y:1986"	"NA8801"	"-"	344975	259221	79682
"Y:1986"	"NA8807"	"X"	344463	259438	80067
"Y:1987"	"NA8801"	"-"	356118	271228	83092
"Y:1987"	"NA8807"	"-"	357583	271924	83732
"Y:1987"	"NA8812"	"X"	357562	272272	83655
"Y:1988"	"NA8901"	"X"	374022	284646	91133

Variable SID (Source IDentification) tells where the data row has come from and when. NA means National Account, and the numbers refer to the year and month of publication of the respective statistics. "Where from" and "when" could be stored separately, but putting them together has created a very informative variable.

OBS and SID are keys. STATUS is an auxiliary variable, whose purpose is to help answer Question 1. Rows with "STATUS" equal to "X" contain the present view of each year's figures. The user interface asks the user, whether he wants only the present view or also older preliminary figures. If the user chooses the present view, he will no longer be bothered with old preliminary figures but gets a time-series composed of all the rows with X-STATUS. These same observations could be sorted out using only the keys, of course, but STATUS makes it much simpler and also faster.

STATUSes are changed when new preliminary figures are published and added to the base. In this respect, STATUS is unique: no other variable's observations are changed afterwards (except to correct typing errors) in the database.

Even though the user may not be interested in old preliminary data, SID contains useful information as it tells when the figures have been updated. Also the user gets the exact sources of the figures: in addition to the table source information (National Accounts) there is a reference to a particular publication for each observation. This information is stored in a separate table, STATINFO, where SID is the key.

Question 2 is easily answered: all figures concerning 1987 are retrieved.

Question 3 can also be answered, but the answer is not ready. The way to do it is to make a new table as follows.

Table 2.

"OBS"	"SID"	"PREL"
"Y:1986"	"NA8710"	"2"
"Y:1986"	"NA8801"	"3"
"Y:1986"	"NA8807"	"4"
"Y:1987"	"NA8801"	"1"
"Y:1987"	"NA8807"	"2"
"Y:1987"	"NA8812"	"3"
"Y:1988"	"NA8901"	"1"

From this new table we can now select e.g. OBS and SID where PREL equals "3", and these OBS and SID combinationss provide the third preliminary figures for the years 1986 and 1987 from table 1.

Notice that table 2 usually cannot be updated mechanically, because the periodicity of the different preliminary figures is not constant. Someone must consider e.g. whether the figures for 1987 that came in December 1988 are similar preliminary figures than those concerning 1986 that came in January 1988. This must be kept in mind when deciding whether this new table is updated continuosly or only when someone wishes to use it.

Although question 3 cannot be answered straight away, the material is in the base and the answer can quite easily be found. Of course, if the publication pattern is highly irregular, the whole question loses its meaning.

Some of the data comes from middle-men who collect it from original sources or from other middle-men's databases. The OECD is an example: it combines data from national sources into its databases and publications. Saving the OECD's old preliminary data does give one a feeling of how this particular middle-man's figures change over time. It usually does not, however, facilitate a thorough study about the properties of the original preliminary figures, because the publication patterns of the OECD and its member countries' statistical offices may not coincide. The obvious way to provide an answer to question 3 in this case would be that the organisation that first stores the data would keep old preliminary figures available for research purposes. In many cases this organisation would be the producer of the original data.

How can we answer question 4? Using SID, we know when the data was available, but only in months. To see if this is sufficient it is necessary to think about the question in more detail.

What we ideally would like to know is what was in the decision-makers head when he made the decision about increasing, for instance, the money supply. What did he know, think and expect? Obviously, this is a bit too much.

Let us be content with what statistical information was available to him. The publication dates of relevant statistics could be stored in the database, and in some cases even hours and minutes. But this is not what he necessarily knew. He may have known less: information systems may still be imperfect. He may as well have known much more: figures published a month later may have already been given to him, which is quite usual at least in Finland. There is no unique time when certain data is first 'available': it is available at different times in different places to different persons.

Probably the best that can systematically and rather effortlessly be done is to store the date when new data was put into the database. It usually is close to the date when the figures were available (barring of course the case when the whole history of a new series is added to the base). This is what can easily and automatically be saved. In ETLA's database the table STATINFO contains an attribute SIDDATE which tells the dates of all the SIDs.

Another possibility would be to add the date after the year and month in the SID. Ordering the data by arrival dates would then, however, be more cumbersome than when using SIDDATEs.

4. Related issues: chained data, forecasts and aggregation in time

There are other figures that bear great resemblance to preliminary data as far as a database structure goes. Prime examples are chained data and forecasts. Also aggregating data in time produces similar figures.

A need to chain data usually arises when base years are changed. Example:

"OBS"	"SID"	"STATUS"	"QQ"	"CQ"	"IQ"
"Y:1970"	"C80B8710"	"X"	205287	151065	58571
"Y:1971"	"C80B8710"	"X"	209575	154918	60780
"Y:1972"	"C80B8710"	"X"	225568	167689	64746
"Y:1973"	"C80B8710"	"X"	240698	177535	70257
"Y:1974"	"C80B8710"	"X"	247987	181752	72710
"Y:1975"	"NA8710"	"X"	250847	188995	77032
"Y:1976"	"NA8710"	"X"	251193	192735	70040

These are again from table 1, some earlier observations. SID now tells that the observations for the years 1970 - 1974 have been

obtained by chaining from the national accounts with 1980 as the base year, and the chaining has taken place in October 1987, when the data was also put into the base. For the years 1975 and 1976, the data was taken from national accounts published in October 1987. This was a case when the whole history was "changed": figures were expressed in 1985 prices instead of earlier 1980 prices.

Forecast data is quite similar:

"OBS"	"SID"	"STATUS"	"QQ_P"	"CQ_P"	"IQ_P"
"Y:1989"	"ETLA8803"	"-"	2.0	3.0	0.5
"Y:1989"	"ETLA8806"	"-"	2.0	3.0	0.5
"Y:1989"	"ETLA8809"	"-"	2.5	3.0	1.0
"Y:1989"	"ETLA8812"	"-"	3.0	3.5	2.0
"Y:1989"	"ETLA8903"	"X"	4.0	3.5	5.5
"Y:1990"	"ETLA8903"	"X"	3.0	3.0	2.5

The variables are percentage change forecasts of production, consumption and investment. The STATUS X now refers to the latest forecast. A table analogous to table 2 is useful to make to facilitate research concerning forecast errors.

Time-aggregation means e.g. calculating quarterly data from monthly data. If original source data is monthly, it may be needed to aggregate it for some purposes into quarterly data. SIDs are useful here too: the SID of the last monthly observation may be used for the quarter, also. It tells that for the other months the observations used in the aggregation were those available at the time of the SID. In practice, we add the letter "A" in front of the new quarterly SID, to remind the user that the data is aggregated. The precise aggregation rules vary between series, of course, and there are many series that cannot be mechanically aggregated.

In some cases it has been necessary to estimate some observation in ETLA or correct the official figures because of some obvious error in official statistics. We have put a special SID for that observation, and kept also the erroneous official figure in the base, for comparison.

5. Conclusions

The costs of the system described above, i.e. the marginal costs of storing old preliminary data also, consists of four groups.

Firstly, planning costs, which were not insignificant but were one-time only. Secondly, updating costs, which are minimal, as they consist of writing the SIDs and STATUSes and the respective sources.

Thirdly come storage costs. Extra observations require disc space. As numerical observations can be packed efficiently and storage capacity has become cheaper, this is not an obstacle.

Fourthly, some costs come in the form of longer query times. Use of STATUS eliminates most of this disadvantage. The decision of whether or not to look at old preliminary figures also takes some time, depending on how it is put into the menu system.

The benefits come also from several sources. Information about preliminary figures is useful, but how useful it is, depends mostly on the user's aims. But it is a benefit not to need to decide separately, what preliminary figures will perhaps be needed in the future, and thus be saved, because this kind of decision making is costly. Multiple users have multiple future needs, impossible to know in advance. ETLA's system has decided it, once and for all: every figure is saved. If one thinks of the alternative costs of collecting preliminary data afterwards, the costs would quite likely be manyfold.

As a by-product, information about sources is more exact than it probably would otherwise be. And also, SID is useful for storing forecasts, chained data, time-aggregated data and all kinds of special observations which in any case would need to be coded exactly.

No exact cost - benefit analysis has been done. We think that the benefits dominate, as is evident from the way we run the data-base.

Appendix: Main features of ETLA's database

ETLA maintains an economic time-series database, the aim of which is to serve ETLA's forecasting activities, model building and various kinds of economic research. The database is also available on-line to outside users on a commercial basis. There are approximately 22 000 time series in the base (1. 5.1989).

The database is built using a Swedish relational database program MIMER. The base is on a Hewlett-Packard 9000 minicomputer. The operating system is HP-UX (unix). PC-users have a terminal emulation program Reflection (by Walker Richer & Quinn, Inc.).

Although a commercial DBMS is used, a lot has been necessary to add to it in order to get a satisfactory system. Especially the user interface and a way to merge data versatilely had to be devised. The user interface is menu-driven and using F-keys. The interface system has been built in ETLA. Combined with the menu system is a transformation program Table, which enables the user to merge data from different database tables and transform it in many ways. Merging requires the program to handle missing data, and Table does this. Table was built in ETLA by Heikki Vajanne, see (5).

Acknowledgements

The author would like to thank Mari Harni, Juha Kinnunen and Heikki Vajanne for helpful suggestions and discussions.

References

(1) Boucelham, J. and T. Teräsvirta (1989). How to use preliminary values in forecasting the monthly index of industrial production? The Research Institute of the Finnish Economy (ETLA), Discussion Paper No 284.

(2) Jong, P. de (1987). Rational economic data revisions. Journal of Business & Economic Statistics 5, 539-548.

(3) Trivellato, U. and E. Rettore (1986). Preliminary data errors and their impact on the forecast error of simultaneous-equation models. Journal of Business & Economic Statistics 4, 445-453.

(4) See almost any issue of International Journal of Forecasting or Journal of Business & Economic Statistics.

(5) Vajanne, H. (1989). Table manual (in Finnish, unpublished).

ELINKEINOELÄMÄN TUTKIMUSLAITOS (ETLA)
The Research Institute of the Finnish Economy
Lönnrotinkatu 4 B, SF-00120 HELSINKI Puh./Tel. (90) 601 322
Telefax (90) 601 753

KESKUSTELUAIHEITA - DISCUSSION PAPERS ISSN 0781-6847

No 276 MIKAEL INGBERG, Näkökohtia metsäverotuksesta. 11.11.1988. 34 s.

No 277 MARKKU KOTILAINEN - TAPIO PEURA, Finland's Exchange Rate Regime and European Integration. 15.12.1988. 37 pp.

No 278 GEORGE F. RAY, The Finnish Economy in the Long Cycles. 20.12.1988. 104 pp.

No 279 PENTTI VARTIA - HENRI J. VARTIAINEN, Finnish Experiences in a Dual Trade Regime. 20.12.1988. 18 pp.

No 280 CHRISTIAN EDGREN, Tulorakenteen hyväksikäytöstä veronalaisen tulon kasvua arvioitaessa. 22.12.1988. 32 s.

No 281 PEKKA ILMAKUNNAS - HANNU TÖRMÄ, Structural Change of Factor Substitution in Finnish Manufacturing. 09.01.1989. 22 pp.

No 282 MARKKU RAHIALA - TIMO TERÄSVIRTA, Labour Hoarding Over the Business Cycle: Testing the Quadratic Adjustment Cost Hypothesis. 18.01.1989. 22 pp.

No 283 ILKKA SUSILUOTO, Helsingin seudun aluetalous panos-tuotostutkimuksen valossa. 08.02.1989. 27 s.

No 284 JAMEL BOUCELHAM - TIMO TERÄSVIRTA, How to Use Preliminary Values in Forecasting the Monthly Index of Industrial Production? 08.03.1989. 14 pp.

No 285 OLLE KRANTZ, Svensk ekonomisk förändring i ett långtidsperspektiv. 28.02.1989. 29 p.

No 286 TOR ERIKSSON - ANTTI SUVANTO - PENTTI VARTIA, Wage Setting in Finland. 20.03.1989. 77 p.

No 287 PEKKA ILMAKUNNAS, Tests of the Efficiency of Some Finnish Macroeconomic Forecasts: An Analysis of Forecast Revisions. 30.03.1989. 19 p.

No 288 PAAVO OKKO, Tuotantomuodon muutos ja sen merkitys yritys- ja aluerakenteelle. 08.05.1989. 14 s.

No 289 ESKO TORSTI, The Forecasting System in ETLA. 10.05.1989. 36 p.

No 290 ESKO TORSTI, MAT-ohjelmointitulkin käyttö ja rakenne. 11.05.1989. 67 s.

No 291 GUJA BACCHILEGA - ROBERTO GOLINELLI, Medium Term Prospects for the European Economies. 17.05.1989. 27 p.

No 292 KARI ALHO, Deregulation of Financial Markets: A General Equilibrium Analysis of Finland. 31.05.1989. 43 p.

No 293 PAAVO OKKO - EERO KASANEN, A Model of Banking Competition. 15.06.1989. 20 p.

No 294 HILKKA TAIMIO, Naisten kotityö ja taloudellinen kasvu Suomessa vuosina 1860-1985. 28.06.1989. 38 s.

No 295 PETTERI HIRVONEN, Kysyntä - tarjonta -kehikon mukainen siirtofunktiomalli bruttokansantuotteelle. 23.08.1989. 38 s.

No 296 PAAVO OKKO, Suomen aluekehityksen ja aluepolitiikan nykyvaihe. 01.09.1989. 20 s.

No 297 ANTTI RIPATTI - PENTTI VARTIA - PEKKA YLÄ-ANTTILA, Suomen talouden ja yritysrakenteen muutokset 1938-1988. 11.09.1989. 95 s.

No 298 ROBERT HAGFORS, On Economic Welfare Equality as a Policy Goal and Social Transfers as Instruments. 11.09.1989. 20 p.

No 299 SYNNÖVE VUORI - PEKKA YLÄ-ANTTILA, Joustava tuotantostrategia puu- ja huonekaluteollisuudessa. 27.09.1989. 60 s.

No 300 SEVERI KEINÄLÄ, Finnish High-Tech Industries and European Integration; Sectoral Study 1: The Telecommunications Equipment Industry. 12.10.1989. 85 p.

No 301 VESA KANNIAINEN, The Arch Model and the Capm: A Note. 30.10.1989. 10 p.

No 302 VESA KANNIAINEN, Research Issues in Corporate Taxation. 30.10.1989. 10 p.

No 303 TOM BERGLUND, Perceived and Measured Risk; An Empirical Analysis. 30.10.1989. 29 p.

No 304 SEVERI KEINÄLÄ, Finnish High-Tech Industries and European Integration; Sectoral Study 2: The Data Processing Equipment Industry. 01.11.1989. 44 p.

No 305 MASSIMO TAZZARI, Numeeriset yleisen tasapainon ulkomaankaupan mallit, teoria ja sovellutukset. 02.11.1989. 64 s.

No 306 JUKKA LASSILA, Preliminary Data in Economic Databases. 10.11.1989. 10 p.

Elinkeinoelämän Tutkimuslaitoksen julkaisemat "Keskusteluaiheet" ovat raportteja alustavista tutkimustuloksista ja väliraportteja tekeillä olevista tutkimuksista. Tässä sarjassa julkaistuja monisteita on rajoitetusti saatavissa ETLAn kirjastosta tai ao. tutkijalta.

Papers in this series are reports on preliminary research results and on studies in progress; they can be obtained, on request, by the author's permission.

0033A/10.11.1989