WORKING PAPER

OUTLINE OF A DATA MODIFICATIONS SYSTEM FOR AGGREGATION OF COMMODITY FLOWS

Alexander Sarkisov
Börje Johansson

November, 1983

WP-83-105

IIASA
International Institute
for Applied Systems Analysis

OUTLINE OF A DATA MODIFICATIONS SYSTEM FOR AGGREGATION OF COMMODITY FLOWS

Alexander Sarkisov
Börje Johansson

November, 1983

WP-83-105

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

FOREWORD

This paper has been motivated by the need to develop software and methods for analyzing, modifying and restructuring data on world trade in the Forest Sector Project. The paper outlines a system of approaches to (i) modifying, correcting, and aggregating initial data on bilateral trade flows and (ii) disaggregating aggregate flows obtained from model projections. The approaches suggested have their origin in experiences with multiregional modeling. The computer programs corresponding to the methods described in the paper have been, and are being, further developed by Alexander Sarkisov who participated in the Regional and Urban Development Group within the YSSP program.

Börje Johansson
Acting Leader
Regional & Urban Development Group

October, 1983

CONTENTS

1. INTRODUCTION 1

1.2 Data Analysis and Data Modification Systems 2

2. AGGREGATION AND DISAGGREGATION OF SPATIAL COMMODITY FLOWS 4

2.1 Specification of Commodity Flows and Composition of Aggregates 4

2.2 Composition of Regional Commodity Groups 6

2.3 Composition of Trade Flows 9

3. HOMOGENEITY OF GROUPS IN AGGREGATION SCHEMES 12

3.1 Direct and Generalized Commodity Homogeneity 12

3.2 Procedures for Creating Commodity Groups 16

3.3 Regional Aggregation with Given Commodity Groups 19

4. AGGREGATION ALGORITHMS: Further Details 22

4.1 An Algorithm for Finding Universal Classifications 22

4.2 An Algorithm for Finding the Permutation Matrix S 24

REFERENCES 26

OUTLINE OF A DATA MODIFICATIONS SYSTEM FOR AGGREGATION OF COMMODITY FLOWS

Alexander Sarkisov
Börje Johansson

1. INTRODUCTION

1.1 Trade Flows, Differentiated Competition and Aggregation

This paper is concerned with problems related to the analysis of trade matrices describing bilateral trade flows and associated price and transportation costs structures. Models for generating scenarios of trade patterns have to be formulated in terms of aggregate trade matrices in order to be computationally tractable. Such aggregation is accomplished by forming groups of commodities (commodity groups) and groups of countries (regions). The aggregate structure should ideally have groups which are as homogenous as possible. This is important for three reasons: (i) to make the model more reliable, (ii) to make it possible to interpret model results in a reasonable way, and (iii) eventually to disaggregate the model results into a finer commodity classification and regional partitioning.

This problem of aggregation-disaggregation is a general problem in trade analysis, since all commodity classifications refer to non-homogenous products and all regions are non-homogenous spatial aggregates. Moreover, the analysis of trade flows brings product differentiation and differentiated competition into focus. According to the classical analysis of Robinson

(1933) and Chamberlin (1933) every bilateral flow should be looked upon as a differentiated commodity, which implies that prices potentially should be considered as link-specific. Another related issue concerns product development which may imply that a commodity group may contain components which change quality and attributes while others remain unchanged over time.

The kind of problems mentioned above are discussed in this paper both from the viewpoint of methods for (i) disaggregating model results and (ii) analyzing historical data in order to create meaningful aggregation schemes.

1.2 Data Analysis and Data Modification Systems

The earlier introduction motivates the need for a data processing system which can be used to modify and restructure a data base. Moreover, such a data base is often created from many different sources which means that the data analysis and data processing must solve problems of consistency. In particular, the reshaping of data should be problem-specific in order to be purposeful. Since the same data base may be used for several different problems, there is a need for a flexible data processing system which can transform the original data to a problem-relevant form (compare Johansson & Marksjö, 1983). This is described in Figure 1.

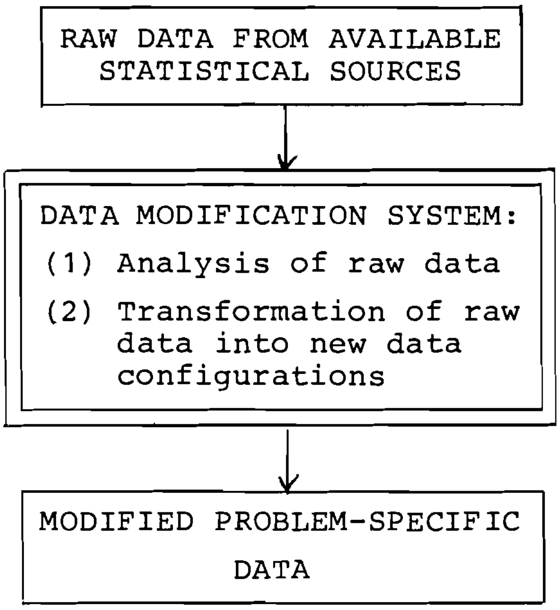
RAW DATA FROM AVAILABLE STATISTICAL SOURCES
DATA MODIFICATION SYSTEM:
(1) Analysis of raw data
(2) Transformation of raw data into new data configurations
MODIFIED PROBLEM-SPECIFIC DATA

Figure 1. Scheme for data transformation.

The analysis in the paper is based on experiences with the Data Modification System which is currently being developed in the Institute for Systems Studies in Moscow (USSR), (Britkov, V., Sarkisov, A., 1983). Such a system may be described as an interactive artificial intelligence system which is based on an ambition to use a language which is close to the natural language related to the given or specified problem. The main operations of such a system are: (i) consistency checking, (ii) classification, (iii) solving for missing information, (iv) aggregation and disaggregation, etc. Those operations may be summarized by the notions data processing and preliminary data analysis.

The presentation also refers to recent results as regards aggregation analysis (Batten, 1982; Roy, Batten & Lesse, 1982; Lesse, Batty & Batten, 1983). Several of the procedures proposed rely on the existence of algorithms of the type described in Eriksson (1981), Andersson & Persson (1982), and Batten (1982). Moreover, the approaches described are related to Swedish experiences of creating flexible multiregional information systems (Johansson & Marksjö, 1983).

The problem analyzed in the paper is illustrated in Figure 2, where x_i^{rs} denotes delivery of commodity i from country r to country s, and where x_I^{RS} has a similar interpretation for aggregates I∋i, R∋r and S∋s

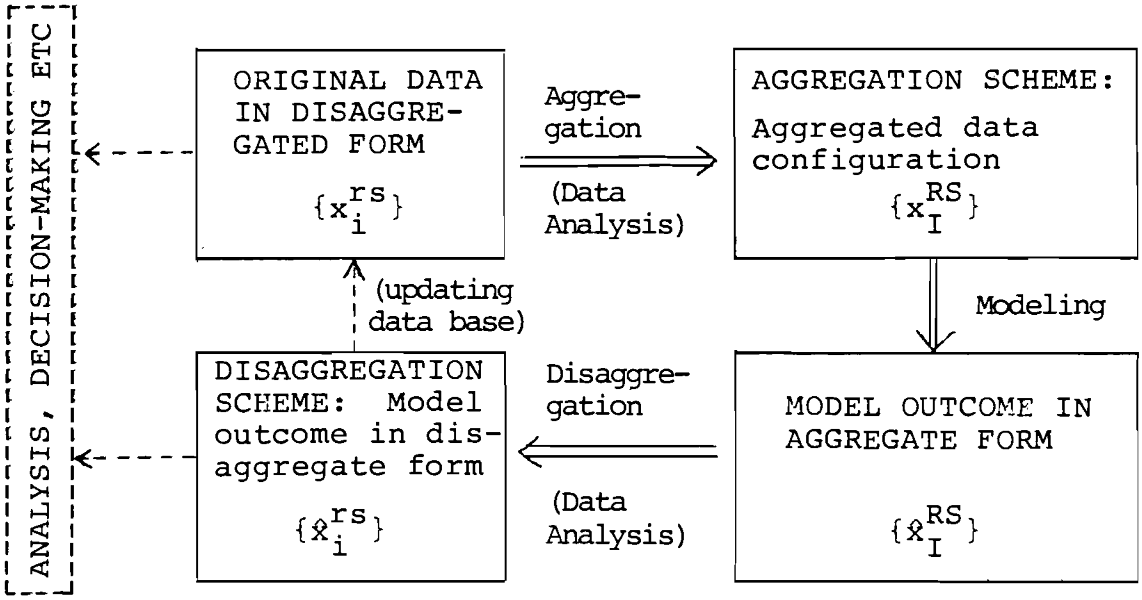
ANALYSIS, DECISION-MAKING ETC
ORIGINAL DATA IN DISAGGREGATED FORM
$\{x_i^{rs}\}$
Aggregation
(Data Analysis)
AGGREGATION SCHEME: Aggregated data configuration
$\{x_I^{RS}\}$
Modeling
MODEL OUTCOME IN AGGREGATE FORM
$\{\hat{x}_I^{RS}\}$
Disaggregation
(Data Analysis)
DISAGGREGATION SCHEME: Model outcome in disaggregate form
$\{\hat{x}_i^{rs}\}$
(updating data base)

Figure 2. Principal procedure for carrying out aggregation-modeling-disaggregation.

2. AGGREGATION AND DISAGGREGATION OF SPATIAL COMMODITY FLOWS

2.1 Specification of Commodity Flows and Composition of Aggregates

We are studying trade flows of commodities between regions. Our analysis refers in particular to models of world trade, in which a region is a group of countries and a commodity is a group of non-identical traded products. A commodity flow is specified in three basic dimensions by:

(a) a *commodity index*, i, signifying to which product class the flow refers

(b) an *origin index*, r, signifying from which region the flow is delivered (exported)

(c) a *destination index*, s, signifying to which region the flow is delivered (i.e., into which region it is imported)

As a point of reference we shall introduce a specification of commodities which we call the most disaggregated. This specification consists of (i) the index set for commodities $i \in \mathbb{I}$, (ii) exporting regions $r \in \mathbb{R}$, and (iii) importing regions $s \in \mathbb{R}$. Subsets of these indices are denoted by $I \subset \mathbb{I}$, $R \subset \mathbb{R}$, and $S \subset \mathbb{R}$. We shall use I, R, and S to denote sets as well as index numbers. In the first case we can write $I \in \{I_1, \ldots, I_n\}$ where $I_i \subset \mathbb{I}$; $R \in \{R_1, \ldots, R_m\}$ where $R_i \subset \mathbb{R}$; and $S \in \{S_1, \ldots., S_k\}$ where $S_i \subset \mathbb{R}$. In the second case we have $I=1,\ldots,n$; $R=1,\ldots,m$; and $S=1,\ldots,k$. With this we define

An aggregation scheme $I=1,\ldots,n$, $r=1,\ldots,m$, and $s=1,\ldots,k$ is complete if simultaneously

$$\begin{aligned} &\cup\{I_j : j=1,\ldots,n\} = \mathbb{I}, \ \cap\{I_j : j=1,\ldots,n\} = \emptyset \\ &\cup\{R_j : j=1,\ldots,m\} = \mathbb{R}, \ \cap\{R_j : j=1,\ldots,n\} = \emptyset \\ &\cup\{S_j : j=1,\ldots,k\} = \mathbb{R}, \ \cap\{S_j : j=1,\ldots,n\} = \emptyset \end{aligned} \tag{2.1}$$

Trade flows between origin, r or R, and destination, s or S, can be measured in quantity terms, signified by x, value terms at the origin, signified by V, value terms at the destination, signified by W. Using the indexation in (2.1), a trade flow can be characterized by $(x_i^{rs},V_i^{rs},W_i^{rs})$ or $(x_I^{RS},V_I^{RS},W_I^{RS})$. Let c_i^{rs} and c_I^{RS} denote unit transportation costs for the links (r,s) and (R,S). Then we can derive an upper value for transportation costs as follows:

$$\begin{aligned} c_i^{rs}\; x_i^{rs} &\le W_i^{rs} - V_i^{rs} \\ c_I^{RS}\; x_I^{RS} &\le W_I^{RS} - V_I^{RS} \end{aligned} \qquad (2.2)$$

Consider now that we have a model with which we are generating projections (scenarios, forecasts, etc.) for aggregate trade flows (I,R,S). We want these flows to represent the individual disaggregate flows which cannot or should not be studied directly. The reason for this may be that the size of a disaggregated model is intractable, and/or that invariances as regards the behavior of the system are more significant in an aggregate form. Our desire is to construct aggregation schemes that make it possible to disaggregate the projections which we obtain from our model. We also want to know which additional information we need to perform the disaggregation. Moreover, even if no disaggregation is carried through, we need conceptual models to interpret aggregate projections, to evaluate uncertainties and pertinent information distortions that are due to the aggregation.

We shall not attempt to provide any definite solutions to the problems mentioned above. Instead we shall examine the possibilities of utilizing a data modification system and associated methods for analyzing data and constructing aggregation schemes which are designed for each specific problem.

With regard to spatial aggregation we may identify the following forms:

$$(x_i^{RS},V_i^{RS},W_i^{RS}) = \sum_{\substack{r\in R\\ s\in S}} (x_i^{rs},V_i^{rs},W_i^{rs}) \qquad (2.3)$$

$$(x_i^{R*},V_i^{R*},W_i^{R*}) = \sum_{\substack{r\in R\\ s\in \mathbb{R}}} (x_i^{rs},V_i^{rs},W_i^{rs}) \qquad (2.4)$$

$$(x_i^{*S}, V_i^{*S}, W_i^{*S}) = \sum_{\substack{s\in S\\ r\in \mathbb{R}}} (x_i^{rs}, V_i^{rs}, W_i^{rs}) \tag{2.5}$$

where x_i^{R*} and x_i^{*S} represent total supply in region R, and total demand in region S. Commodity aggregation takes the form

$$(x_I^{rs}, V_I^{rs}, W_I^{rs}) = \sum_{i\in I} (x_i^{rs}, V_i^{rs}, W_i^{rs}) \tag{2.6}$$

The remaining forms of aggregation are obtained by combining (2.6) with (2.3)-(2.5).

2.2 Composition of Regional Commodity Groups

The most disaggregate form of regional commodities as given by (2.4) and (2.5) is $x_i^{r*} = \Sigma_s x_i^{rs}$ and $x_i^{*s} = \Sigma_r x_i^{rs}$, where the first denotes regional supply (at origin r) and the second regional demand (at destination s). We should observe that in practice x_i^{r*} is an aggregate summarizing the supply of different firms whose products have non-identical attributes; they may differ with regard to technical characteristics, but also in terms of logistics, guarantee of quality provided by a good name, quickness of service, length of credit, advertisement, etc., (compare Robinson, 1933). These aspects become even more evident when we are forming commodity aggregates

$$x_I^{r*} = \sum_{i\in I} x_i^{r*} \tag{2.7}$$

$$x_I^{*s} = \sum_{i\in I} x_i^{*s} \tag{2.8}$$

Using the same type of summation as in (2.7) and (2.8) we can form the price variables $v_I^{k*} = V_I^{k*}/x_I^{r*}$ and $w_I^{*k} = W_I^{*k}/x_I^{*k}$. If there is no price dispersion and competition is perfect the regional price, p_I^r, satisfies the "ideal" condition [1]

$$p_I^k = v_I^{k*} = w_I^{*k} \tag{2.9}$$

[1] In order to avoid a jungle of notations we only consider (throughout the paper) prices which are obtained after the effects of tariffs and the like have been removed.

In order to draw conclusions about regional prices of individual products, p_i^r, from observations of p_I^r, we must use some assumption about the aggregate commodity I. One may, e.g., assume that I is a composite commodity, which has the following meaning

> Let t denote time. Then a commodity group I is a *composite commodity* if for every $i \in I$ there exists a positive time-invariant scalar α_j^r such that we have $p_j^r(t) = \alpha_j^r \, p_I^r(t)$. (2.10)

Suppose that products are developed over time so that the attributes of commodity $j \in I$ are changed in such a way that the quality is improved. This might give rise to a change process in which $\alpha_j^r(t)$ increases over time. For another commodity $i \in I$ we may simultaneously observe that $\alpha_i^r(t)$ gradually decreases over time. From this we may define

> The commodity group I is a *composite commodity in a weak sense* if for each $i \in I$ there exist *known* functions $\alpha_i^r(t) > 0$ such that we have $p_i^r(t) = \alpha_i^r(t)$ $p_I^r(t)$. (2.11)

Suppose that we have information about the two aggregates v_I^{r*} and x_I^{r*} and want to decompose this information into the disaggregated elements v_i^{r*} and x_i^{r*}, where v_I^{r*} may be regarded as a proxy for p_I^r. In order to carry through the decomposition we need the following strong condition

> A commodity group I is *regionally decomposable* if for each $i \in I$ there exists a pair of known functions $(\alpha_i^r(t), \beta_i^r(t))$ such that (2.12)
>
> (i) $p_i^r(t) = \alpha_i^r(t) \; p_I^r(t)$,
>
> (ii) $x_i^{r*}(t) = \beta_i^r(t) \; x_I^{r*}(t)$,
>
> (iii) $\sum_i \alpha_i^r(t) \beta_i^r(t) = 1$ and $\sum_i \beta_i^r(t) = 1$.

where condition (iii) is a result of the simultaneous fulfilment of (i) and (ii).

Let $V_I^{r*}(t)$ and $x_I^{r*}(t)$ be forecasted values. It means that the regional price of commodity I is known, $v_I^{r*}(t) = V_I^{r*}(t)/x_I^{r*}(t)$, and can serve as a proxy for $p_I^r(t)$. Consider then the case when we assume I to be a composite commodity in the weak sense ($\alpha_i^r(t)$ are known functions of time). If for this case we also have information (or introduce assumptions) about the maximum supply capacity for individual products $X_i^r(t)$ then the fact that values $\beta_i^r(t) \geq 0$ and $\sum_i \beta_i^r = 1$ and thus can be considered as probability distributions, allows us to use the information theory approach to calculate the most probable values of $\beta_i^r(t) = \hat{\beta}_i^r(t)$ by solving for any fixed time point the following minimization problem:[1]

$$\begin{aligned} &\text{Min } \sum_i \beta_i^r \log \beta_i^r \\ &\text{s.t.} \\ &\beta_i^r x_I^r \leq X_i^r \\ &\sum_i \alpha_i^r \beta_i^r = 1 \end{aligned} \qquad (2.13)$$

Observe that in this case we know cost functions of individual products $p_i^r(t) = \alpha_i^r(t)\ p_I^r(t)$.

Now for the same forecasted values consider the case when the sectoral change paths $\beta_i^r(t)$ and additional information about maximum price values for individual products $p_i^r(\max)$ are given. If we substitute new variables $\xi_i^r(t)$ for α_i^{rs} so that $\xi_i^r(t) = \alpha_i^r(t)\ \beta_i^r(t)$ then the condition for regional decomposition of commodity I can be rewritten:

[1] The solution to (2.13) could also be called "maximally non-committed with regard to missing information" (Jaynes, 1957).

$$\text{(i)} \quad p_i^r(t) = (\xi_i^r(t)/\beta_i^r(t))\ p_I^r(t)$$

$$\text{(ii)} \quad x_i^{r*}(t) = \beta_i^r(t)\ x_I^r(t) \qquad (2.12')$$

$$\text{(iii)} \quad \sum_i \xi_i(t) = 1 \quad \text{and} \quad \sum_i \beta_i^r(t) = 1$$

The fact that $\xi_i(t) = \alpha_i(t)\ \beta_i(t) \geq 0$ and $\sum_i \xi_i(t) = 1$ together with the above mentioned assumptions again allows us to use the information theory methodology to calculate the most probable values $\xi_i(t) = \hat{\xi}_i(t)$ (and thus $\alpha_i(t) = \hat{\alpha}_i(t)$) by solving at any fixed time point the following minimization problem:

$$\text{Min} \sum_i \xi_i^r \log \xi_i^r$$

$$\text{s.t.} \qquad (2.14)$$

$$\xi_i^r\ p_I^r \leq \beta_i^r(t)\ p_i^r(\text{max})$$

2.3 Composition of Trade Flows

Consider the trade flows originating from region r. As far as quality requirements and demand for attributes differ in two specific destination regions, s and k, we should in general expect x_I^{rs} to have a composition different from x_I^{rk}.[2] Such a differentiation of flows is a natural extension of the heterogeneity of regional commodities x_I^{r*} as described in section 2.2.

[2] The difference in composition means, e.g., that $\beta_i^{rs} \neq \beta_i^{rk}$, $i \in I$.

Referring to formulae (2.2), (2.3) and (2.9) we shall assume that W_i^{rs} and V_i^{rs} are measured in such a way that flow prices, p_i^{rs}, can be calculated as $p_i^{rs} = V_i^{rs}/x_i^{rs} = W_i^{rs}/x_i^{rs} - c_i^{rs}$. Using the notation $w_i^{rs} = W_i^{rs}/x_i^{rs}$ we shall examine the following four types of flow prices:

$$p_i^{rs} = w_i^{rs} - c_i^{rs}$$

$$p_I^{rs} = \sum_{i\in I} (w_i^{rs}-c_i^{rs})x_i^{rs}/x_I^{rs} = w_I^{rs} - c_I^{rs}$$

$$p_i^{RS} = \sum_{r\in R} \sum_{s\in S} (w_i^{rs}-c_i^{rs})x_i^{rs}/x_i^{RS} = w_i^{RS} - c_i^{RS} \tag{2.15}$$

$$p_I^{RS} = \sum_r \sum_s (w_I^{rs}-c_I^{rs})x_I^{rs}/x_I^{RS} =$$

$$= \sum_{i\in I} (w_i^{RS}-c_i^{RS})x_i^{RS}/x_I^{RS} = w_I^{RS} - c_I^{RS}$$

In Figure 3 we describe some of the possible disaggregation schemes that may be applied to an aggregate flow (I,R,S) characterized by p_I^{RS} and x_I^{RS}. The boxes in the figure with double borderlines indicate those levels of disaggregation on which our interest will be concentrated.

Let us illustrate the process of disaggregation by examining the following two cases:

(i) $\{p_I^{RS},x_I^{RS}\} \to \{p_i^{RS},x_i^{RS}\}$, and

(ii) $\{p_I^{RS},x_I^{RS}\} \to \{p_I^{rs},x_I^{rs}\}$,

where $\{p_I^{RS},x_I^{RS}\}$ represents the outcome of a model projection for a future point in time.

Suppose that additional information which has not been utilized in the model projection is available. With regard to case (i), let this information be

- An a priori probability distribution $\bar{\beta}_i^{RS}$, indicating the expected likelihood that an arbitrary element in the flow x_I^{RS} belongs to commodity category i.

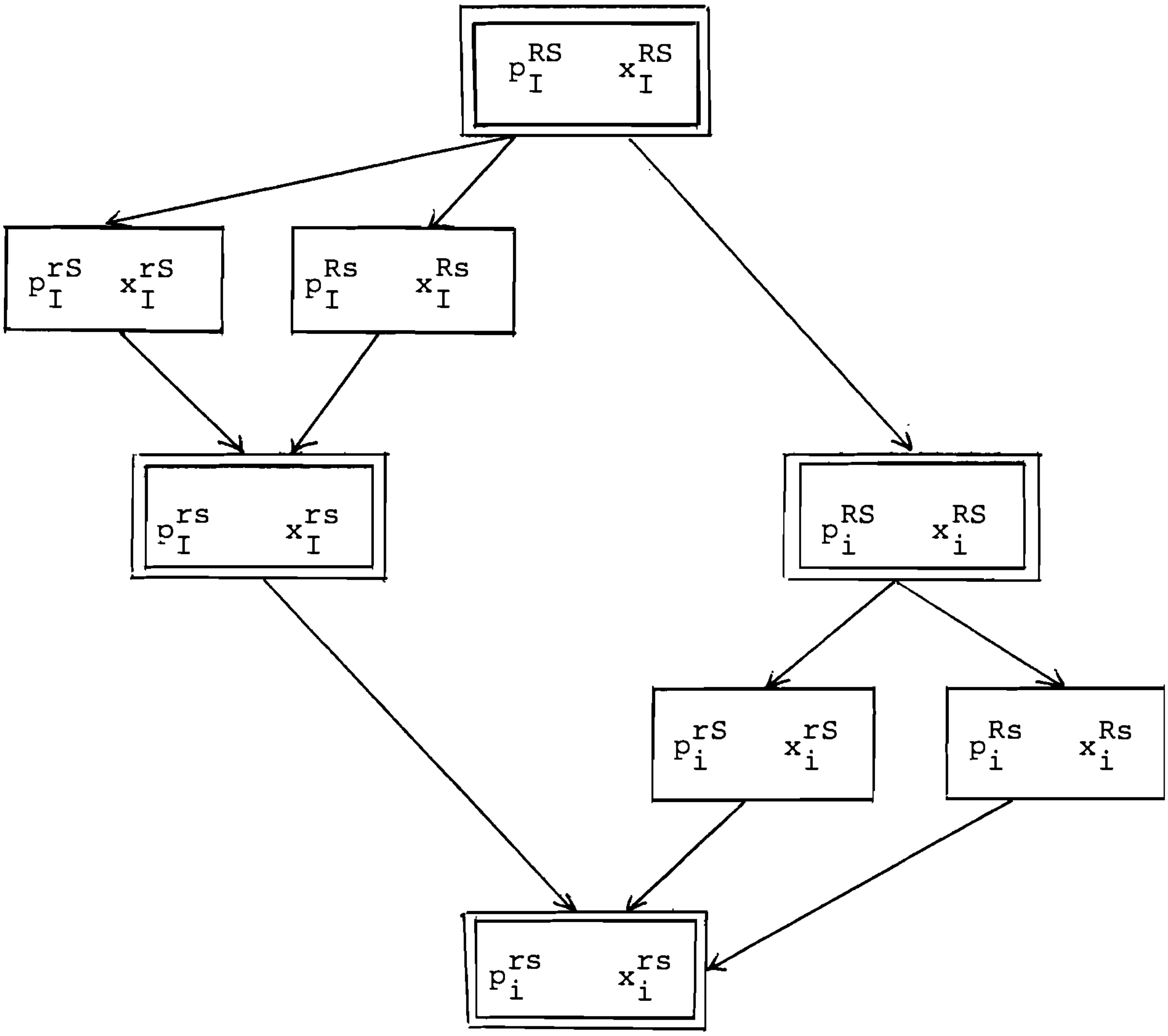

Figure 3. Illustration of disaggregation schemes.

- Estimates of transportation costs c_i^{RS}.
- Price scenarios $p_i^{RS} = \alpha_i^{RS} p_I^{RS}$ for individual commodity flows x_i^{RS}.

Given this information, we may apply a minimum information principle to calculate the most probable values $\beta_i^{RS} =$ $= x_i^{RS}/x_I^{RS}$.[1] This procedure may be summarized by the following Lagrange function

$$L = \sum_i \beta_i^{RS} \log \beta_i^{RS}/\bar{\beta}_i^{RS} + \lambda \sum \beta_i^{RS}(1-\alpha_i^{RS}) + \gamma[\sum \beta_i^{RS} c_i^{RS} - c_I^{RS}] \qquad (2.16)$$

[1] Compare Snickars and Weibull (1977).

This yields the solution

$$\beta_i^{RS} = \frac{\bar{\beta}_i^{RS} \exp\{\lambda \alpha_i^{RS} - \gamma c_i^{RS}\}}{\sum_i \bar{\beta}_i^{RS} \exp\{\lambda \alpha_i^{RS} - \gamma c_i^{RS}\}} \tag{2.17}$$

An analogous approach with regard to case (ii) yields

$$\beta_I^{rs} = \frac{\bar{\beta}_I^{rs} \exp\{\lambda \alpha_I^{rs} - \gamma c_I^{rs}\}}{\sum_{rs}\sum \bar{\beta}_I^{rs} \exp\{\lambda \alpha_I^{rs} - \gamma c_I^{rs}\}} \tag{2.18}$$

Formulae (2.16) and (2.17) illustrate how alternative price scenarios affect the β_I^{rs}- and β_i^{RS}-distributions which are compatible with a given aggregate projection.

3. HOMOGENEITY OF GROUPS IN AGGREGATION SCHEMES

An aggregation scheme consists of groups formed with regard to characteristics in three dimensions (i,r,s) where $i \in II$ and $r,s, \in \mathbb{R}$. As described in (2.1) a group is identified by a triple (I,R,S). A purposeful aggregation has to be constructed on the basis of a set of criteria, which is selected contingent on the modeling purpose. Here we shall confine the discussion to criteria which require that each group is homogenous in a problem-specific sense. Since we are considering projections into the future of aggregate flows x_I^{RS}, we want the homogeneity property of the elements x_i^{rs}, where $(i,r,s) \in (I,R,S)$, to be approximately time invariant.

3.1 Direct and Generalized Commodity Homogeneity

In section 2.2 regional commodity groups were introduced. From (2.7) and (2.8) we can define

$$\begin{aligned} x_I^{R*} &= \sum_S x_I^{RS} = \sum_{i \in I} x_i^{R*} \\ x_I^{*S} &= \sum_R x_I^{RS} = \sum_{i \in I} x_i^{*S} \end{aligned} \tag{3.1}$$

Homogeneity may be seen as a key to disaggregation of observed or known aggregates. For a regional supply group (I,R,*) we may consider the following criterion:

> *Input homogeneity:* this type of homogeneity is satisfied if the input coefficient vector of every good $i \in I$ is proportional to the aggregate input vector referring to I as a whole. (3.2)

Condition (3.2) implies that cost conditions are similar for all $i \in I$. It is easy to see that this condition only can be fulfilled if either (i) input coefficients are constant or (ii) production levels are proportional so that every ratio x_i^{R*}/x_I^{R*} is constant.

For a regional demand group (I,*,S) we may consider two types of criteria. The first requires that all $i \in I$ are "decomposable substitutes" in the sense specified below. The other requires that the goods $i \in I$ are "decomposable complements".

> *Observation 1:* Suppose that all $i \in I$ are perfect substitutes in the sense that there is only one market price w_I^{*S}. Then the goods are decomposable substitutes if there exist separable functions g_i^S such that $x_i^{*S} = g_i^S(w_I^{*S})$ for each $i \in I$. If this is fulfilled all information about the disaggregation scheme $\{w_i^{*S}, x_i^{*S}\}$ is contained in w_I^{*S}.

> *Observation 2:* Suppose that all goods $i \in I$ are perfect complements in that there exist coefficients such that $\delta_i^{*S} = x_i^{*S}/x_I^{*S}$. These different goods are decomposable complements if there exist separable functions f_i^S such that $w_i^{*S} = f_i^S(x_i^{*S})$. If this is satisfied all information about $\{x_i^{*S}, w_i^{*S}\}$ is contained in x_I^{*S}.

Consider now the problem of aggregating commodities to a group I in such a way that the flow (R,S) is decomposable.

An extension of (2.12) yields the following criterion for *strong flow decomposability:*

$$\begin{aligned} x_i^{RS}(t) &= \beta_i^{RS} x_I^{RS}(t) \quad \text{for each } i \in I \\ p_i^{RS}(t) &= \alpha_i^{RS} p_I^{RS}(t) \quad \text{for each } i \in I \end{aligned} \tag{3.3}$$

In this case the commodity composition and the relative prices are simultaneously time invariant, and the aggregate pair $(x_I^{RS}(t), p_I^{RS}(t)$ is a perfect proxy for the individual flow terms $\{p_i^{RS}(t), x_i^{RS}(t)\}$.

> *Observation 3:* Time invariance of coefficients α_i^{RS} is not a sufficient condition to guarantee time invariant price ratios $w_i^{RS}(t)/w_I^{RS}(t)$. However, strong flow decomposability ensures this latter time invariance.

This observation follows directly from the definition $w_i^{RS}(t) = \alpha_i^{RS} p_I^{RS}(t) + c_i^{RS}$ and $w_I^{RS}(t) = p_I^{RS}(t) + \Sigma c_i^{RS} \beta_i^{RS}(t)$.

Suppose we have a time series $T = \{1,\dots,\bar{T}\}$ and wish to investigate the degree of invariances of the type described in (3.3). This could be done in hierarchical steps, starting with calculating correlation coefficients ρ_{ij}^{β} and ρ_{ij}^{α} where for a given complete spatial partitioning $R \subset \mathbb{R}$ and $S \subset \mathbb{R}$.

$$\begin{aligned} \rho_{ij}^{\beta} &= \underset{\substack{\{R,S\} \\ t \in T}}{\text{Corr}} \left(x_i^{RS}(t), x_j^{RS}(t)\right) \\ \rho_{ij}^{\alpha} &= \underset{\substack{\{R,S\} \\ t \in T}}{\text{Corr}} \left(p_i^{RS}(t), p_j^{RS}(t)\right) \end{aligned} \tag{3.4}$$

We may split the correlation measures in (3.4) into (i) time-specific measures $(\rho_{ij}^{\beta}(t), \rho_{ij}^{\alpha}(t))$ and (ii) flow specific measures $(\rho_{ij}^{\beta(R,S)}, \rho_{ij}^{\alpha(R,S)})$ such that

$$\rho_{ij}^{\beta}(t) = \underset{\{R,S\}}{\mathrm{Corr}}\,(x_i^{RS}(t), x_j^{RS}(t)) \quad \text{for } t \text{ given}$$

$$\rho_{ij}^{\beta(R,S)} = \underset{t\in T}{\mathrm{Corr}}(x_i^{RS}(t), x_j^{RS}(t)) \text{ for } (R,S) \text{ given}$$

and analogously for $\rho_{ij}^{\alpha}(t)$ and $\rho_{ij}^{\alpha(R,S)}$.

Observation 4: Let $i,j\in I$. If α_i^{RS} and α_j^{RS} are time invariant, $\rho_{ij}^{\alpha(R,S)}$ will be close to one. If α_i^{RS} and α_j^{RS} are time invariant for each pair R,S in a complete spatial partitioning of $\mathbb{R} \times \mathbb{R}$, ρ_{ij}^{α} will be close to one. Observe finally that $\rho_{ij}^{\alpha(R,S)}$ close to unity does not imply time invariance of α_i^{RS} and α_j^{RS}.

The first part of Observation 4 is self-evident. The second part is obvious if we note that $p_i^{RS}(t)$ and $p_j^{RS}(t)$ are perfectly correlated if $p_i^{RS}(t) = a_i p_I^{RS}(t) + b$ and $p_j^{RS}(t) = = a_j p_I^{RS}(t) + b$.[1] However, $\alpha_i^{RS}(t) = a_i + b/p_I^{RS}(t)$ will be time invariant only if $p_I^{RS}(t)$ is constant.

Observation 5: The conclusions in Observation 4 are also true when β is substituted for α.

Observation 6: Suppose that there is a time invariant relation between $\beta_i^{RS}(t)$ and $\beta_j^{RS}(t)$ such that $\beta_i^{RS}(t) = b_{ij}\beta_j^{RS}(t)$ holds approximately for b_{ij} fixed. Then $\rho_{ij}^{\beta(R,S)}$ will be close to unity. Similarly, $\rho_{ij}^{\alpha(R,S)}$ will be close to unity if $\alpha_i^{RS}(t) = a_{ij}\alpha_j^{RS}(t)$ holds approximately over time for a_{ij} fixed.

The conclusions in Observation 6 are obvious, since if $x_i^{RS}(t) = \beta_i^{RS}(t)x_I^{RS}(t)$ for each $i\in I$, then $x_j^{RS}(t)/x_i^{RS}(t) = 1/b_{ij}$ if $i,j\in I$.

[1] This type of relation can only be valid when p_I^{RS} is constrained to satisfy $p_I^{RS} \geq \pi$ for some $\pi > 0$.

Obviously we can also establish other forms of correlation and time-invariant properties. For example, if $\beta_i^{RS}(t) = \bar{\beta}_i^{RS} f_i^{RS}(t)$ and $\beta_i^{RS}(t) = \bar{\beta}_j^{RS} f_j^{RS}(t)$, then we can examine the correlation between $f_i^{RS}(t)$ and $f_j^{RS}(t)$, which is perfect if, for example, $f_i^{RS}(t) = 1/f_j^{RS}(t)$.

We may remark that the property of flow homogeneity has been transformed to find commodity groups within which there is a strong correlation between β_i^{RS}- and α_i^{RS}-coefficients or time patterns of such coefficients, respectively.

3.2 Procedures for Creating Commodity Groups

We shall describe a step-wise hierarchical procedure which starts from the fine level of commodity classification. In a first step each link (r,s) is examined.[1] For each such link commodities are combined into commodity groups $\{I^1(r,s),...,I^m(r,s)\} = I(r,s)$. Such a constellation of commodity groups is called a commodity classification.

In the subsequent step exporting regions r which have "similar" classifications are grouped together which generates new link pairs (R,s). Each such pair contains a set of classifications $I(R,s) = \{I(r,s):r \in R\}$ which are similar but not necessarily identical. The final step of the exporter-based aggregation consists of finding for each pair (R,s) one single classification $I^*(R,s)$ which can represent the set $I(R,s)$ in a satisfactory way.

The three steps consist of finding (i) $I(r,s)$, (ii) $I(R,s)$, and (iii) $I^*(R,s)$. Before this a correlation matrix $\{\rho_{ij}^{rs}\}$ is calculated for each given pair (r,s) and expressing the correlation between all possible pairs $i,j \in II$. Referring to (3.6) we can calculate two versions of this matrix, namely $\rho_{ij}^{\beta(r,s)}$ and $\rho_{ij}^{\alpha(r,s)}$.[2]

The objective is now to utilise the information in the correlation matrix for a given pair (r,s) in order to form groups containing commodities which strongly correlate with each

[1] This procedure may also start with already spatially aggregated links (R,S) as in Secton 3.1.

[2] Naturally it is also possible to calculate similar correlation coefficients with regard to transportation costs.

other. The following procedure will be outlined. We are searching for a threshold value ρ^* such that all pairs which simultaneously have pairwise values $\rho_{ij} \geq \rho^*$ are included in the same group. Selecting $\rho^* = 1$ means that the number of groups equals the total number of commodities. Reciprocally, selecting $\rho^* = 0$ allows all commodities to be clustered together in one single group. In Figure 4 we have illustrated how the number of groups in a classification varies with the selection of a threshold value. The depicted curve makes a "jump" around the critical value ρ^{**}, i.e., in the neighbourhood of ρ^{**} curve $N(\rho^*)$ has a significantly steeper slope than elsewhere; $N(\rho^*)$ expresses the number of groups obtained when ρ^* is selected.

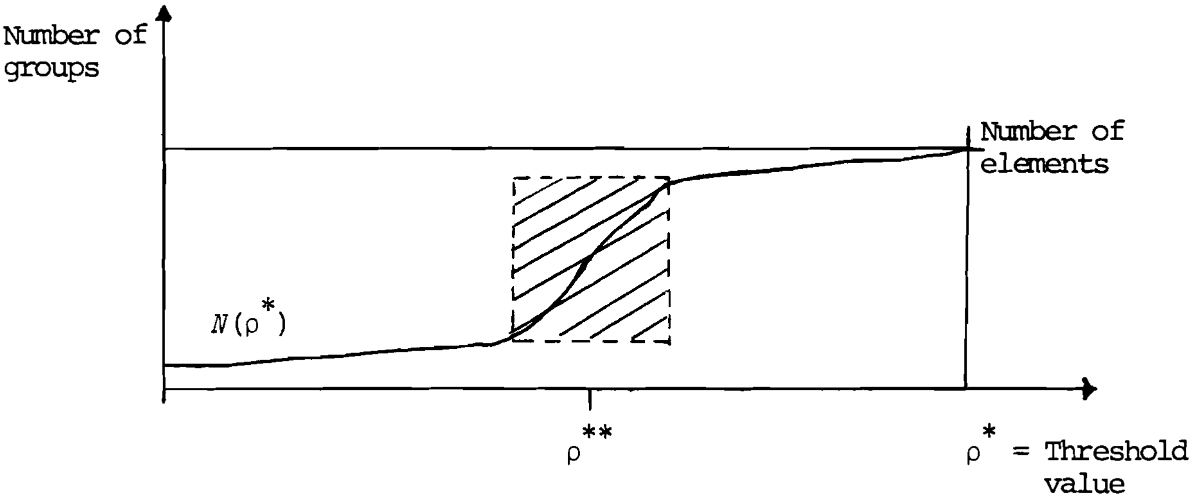

Figure 4. Number of groups in a classification as a function of ρ*.

A desirable solution to the approach described in Figure 4 is obtained if we can find ρ^{**} which is close to one and which at the same time corresponds to a small number of groups in the classification. If the function $N(\rho^*)$ increases without "jumps", for example, when $N(\rho^*)$ has an almost constant first derivative, an exogenously determined compromise between the value of $N(\rho^*)$ and ρ^* has to be utilized.

Suppose that ρ^{**} has been selected. Then the following approach may be applied:[1]

[1] Instead of this heuristic approach one may consider techniques like factor analysis.

- o Commodity i is compared with all $j \neq i$, and we select all j such that $\rho_{ij}^{rs} \geq \rho^*$; these j's form a group I together with i.
- o Another commodity $k \notin I$ is selected and the process is repeated.

With this procedure we have obtained one commodity classification $I(r,s)$ for each pair (r,s), $r,s \in \mathbb{R}$.

The next step consists in bringing links (r,s) and (k,s) together in an export-related aggregate (R,s); the criterion for such spatial "clustering" of export regions (for given import region, s) is that the classifications $I(k,s)$ and $I(r,s)$ are similar for $r,k \in R$. In order to obtain a measure of similarity we first define

$$\theta_{ij}^{ks} = \begin{cases} 1 \text{ if commodity i and j belong to the same group according to classification } I(k,s) \\ 0 \text{ otherwise} \end{cases} \qquad (3.7)$$

Hence, $\theta_{ij}^{ks} = 1$ if $i,j \in I$ and $I \in I(k,s)$. With this auxiliary variable we may calculate the distance between two classifications $I(k,s)$ and $I(r,s)$ as follows

$$d[I(k,s),I(r,s)] = \sum_{i,j \in II} |\theta_{ij}^{ks} - \theta_{ij}^{rs}| \qquad (3.8)$$

Once these distances have been determined we may apply the same technique as was outlined for the correlation coefficients, and illustrated in Figure 4. The outcome will be (R,s)-specific commodity classifications as illustrated in Figure 5.

Our next task is to further compare the individual classifications $I(r,s) \in I(R,s)$, where $r \in R$. On the basis of such a comparison we shall construct for each pair (R,s) a universal classification $I^*(R,s)$ which will be used as a representative classification for (R,s). This is accomplished by solving the following optimization problem:[1]

[1] The algorithm for solving this problem is outlined in Section 4.1.

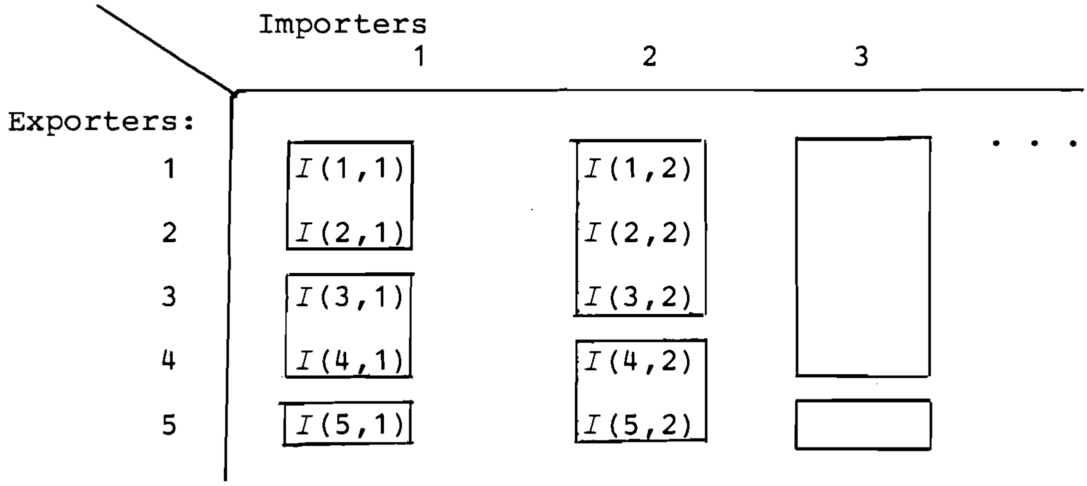

Figure 5. Illustration of $I(R,s)$ and its elements $I(r,s)$ for each importer s.

$$\underset{\{I^*(R,s)\}}{\text{Min}}\ D\ ;\quad \sum_{r\in R} d[I^*(R,s),I(r,s)] = D \tag{3.9}$$

When this is solved we have obtained one classification for each pair (R,s).

At this stage we may identify a final aggregation problem which consists in forming groups of importers into importing regions. The outcome is pairs (R,S) and corresponding universal commodity classifications $I^*(R,S)$, such that individual importers $s \in S$ are similar both with regard to exporter linkages and commodity classifications.

It should be observed that the hierarchical procedure makes it possible to collect structural information at each step of the procedure. Note also, that the aggregation may utilize more than one characteristic per link. Two such characteristics are the α- and β-coefficients as defined in (2.12) and (3.3). Still one more characteristic is the transportation coefficients c_i^{rs} introduced in (2.2).

3.3 Regional Aggregation with Given Commodity Groups

Consider the case of a given commodity i, or a given commodity group I. In the latter case we may, for example, perceive I as a group formed with all i and j such that each

$$\rho^{\alpha}_{ij} = \underset{\substack{r,s \in \mathbb{R} \\ t \in T}}{\mathrm{Corr}} (p^{rs}_{i}(t), p^{rs}_{j}(t))$$

has a high value, where ρ^{α}_{ij} is the price correlation coefficient defined in (3.4).

In this subsection we restrict our analysis to one given commodity group. Therefore we can suppress or delete the commodity index. Our objective is to group together regions which simultaneously have similar export and import patterns as regards the given commodity. Each region may then be considered as a composite exporter and importer.

Trade patterns may be characterized as structural phenomena and reformulated as bivariate probability distributions for which information theory concepts are applicable (e.g., Andersson & Persson, 1982; Batten, 1982). For a given commodity group, this type of probability distribution is comprised of a matrix $Q = \{q^{rs}\}$ where

$$q^{rs} = x^{rs} / \sum_{r,s \in \mathbb{R}} x^{rs} \qquad (3.7)$$

and where x^{rs} denotes the flow between country r and s. The degree of dispersion or uncertainty in such a distribution may be calculated as $\Sigma_{rs} q^{rs} \log q^{rs}$. Transforming the matrix Q to another bivariate probability matrix Z then gives rise to the following information loss between Q and Z or change in the dispersion structure (see e.g., Snickars & Weibull, 1977): [1)]

$$I(Q:Z) = \sum_{r,s \in \mathbb{R}} q^{rs} \log (q^{rs}/z^{rs}) \qquad (3.8)$$

This type of measure will play a central role in the subsequent aggregation procedure. The aggregation of flows (r,s) to flows (R,S) gives us new probabilities

$$q^{RS} = \sum_{r \in R} \sum_{s \in S} q^{rs} \quad , \quad R,S \subset \mathbb{R} \qquad (3.9)$$

[1)] This may be compared with the procedure in (2.16). Observe that I(Q:Z) may also be interpreted as the information gain between Z and Q.

Let $\bar{Z} = \{\bar{z}^{RS}\}$ be an m x m-matrix and let the initial probability array Q be an n x n-matrix with m < n. Then we can find a permutation matrix n x m such that [1]

$$\bar{Z} = S^T QS \tag{3.10}$$

Every row of S has one unit element with all other elements being zeros. The non-zeros in each column show which elements are united into regional groups.

Consider now an auxiliary matrix Z with dimension n x n, which is "information equivalent" with $\bar{Z}$. We may define Z as a matrix obtained by

$$\min_{\{z^{rs}\}} \Sigma z^{rs} \log z^{rs} \tag{3.11}$$

subject to summation constraints with regard to terms in Z which are associated with terms in $\bar{Z}$. In order to find an aggregate matrix $\bar{Z}$ which has a "similar" structure to Q, we solve the following optimization problem (recalling the relation between Z and $\bar{Z}$):

$$\text{Min } I(Q:Z) = \text{Min } \Sigma\, q^{rs} \log (q^{rs}/z^{rs}) \tag{3.8}$$

where I(Q:Z) expresses the information loss between Q and Z caused by choosing Z instead of Q. As shown in Roy et al (1982) we can rewrite I(Q,Z) as follows:

$$I(Q:Z) = \sum_{r=1}^{n} \sum_{s=1}^{n} q^{rs} \log q^{rs} - \sum_{I=1}^{m} \sum_{J=1}^{m} \bar{z}_{IJ} \log (\bar{z}_{IJ}/d(I)d(J)) \tag{3.8'}$$

where n and m are the number of rows (and columns) in the initial and aggregate matrices respectively, and where d(I) is the I'th element in the diagonal matrix D such that

$$D = S^T S \tag{3.12}$$

[1] S^T denotes the transpose of S.

In this way I(Q:Z) is given as a direct function of the permutation matrix S. According to (3.10) this matrix is all we need in order to perform the aggregation.

> *Observation 7:* Suppose that we have obtained aggregate probabilities (share coefficients) $\bar{z}_I^{RS}$. Then
>
> (i) $\sum_{r,s} z_i^{rs}/\bar{z}_I^{RS}$ will represent β_i^{RS} as described in (2.17) and (3.3)
>
> (ii) $\sum_{i \in I} z_i^{rs}/\bar{z}_I^{RS}$ will represent β_I^{rs} as described in (2.18).

4. AGGREGATION ALGORITHMS: Further Details

In section 3.2 we described a procedure for aggregating commodities into classifications of commodity groups, aggregating exporter and/or importer regions, and identifying universal commodity classifications in the following steps

$$\{\rho_{ij}^{rs}\} \rightarrow \mathcal{I}(r,s) \rightarrow \mathcal{I}(R,s) \rightarrow \mathcal{I}^*(R,s)$$

If we like, this sequence can be continued further to establish a globally universal commodity classification $\mathcal{I}^*(*,*)$, for which we can carry out a regional aggregation as described in section 3.3.

As an alternative, the procedure may start with a regional aggregation over a whole set of commodities, and thereafter construct $\mathcal{I}(R,S)$, $\mathcal{I}^*(R,S)$, and eventually $\mathcal{I}^*(*,*)$. Obviously, the search for universal classifications of different orders constitute a fundamental part of the aggregation.

4.1 An Algorithm for Finding Universal Classifications

For each pair (R,s) or (R,S) we are given a classification $\mathcal{I}(r,s)$ or $\mathcal{I}(r,S)$ for each r∈R. The set of such classifications is $\mathcal{I}(R,s)$ or $\mathcal{I}(R,S)$. With regard to the pair (R,s) our objective is to find one classification $\mathcal{I}^*(R,s)$ that can universally represent all classifications in $I(\mathrm{R,s})$. This problem is posed in (3.9). Since the variables θ_{ij}^{rs} in (3.9) are either unity or zero, the expression for D can be modified in the following way:

$$D = \sum_r \sum_{i,j} |\theta_{ij}^{*s} - \theta_{ij}^{rs}| = \sum_r \sum_{i,j} (\theta_{ij}^{*s} - \theta_{ij}^{rs})^2 =$$

$$= \sum_r \sum_{i,j} \theta_{ij}^{*s} - 2 \sum_r \sum_{i,j} \theta_{ij}^{*s}\theta_{ij}^{rs} + \sum_r \sum_{i,j} \theta_{ij}^{rs} = \qquad (4.1)$$

$$= N \sum_{i,j} \theta_{ij}^{*s} - 2 \sum_{i,j} \theta_{ij}^{*s}[\sum_r \theta_{ij}^{rs}] + \sum_{i,j} \sum_r \theta_{ij}^{rs}$$

Let us introduce the variable $a_{ij} = \Sigma_r \theta_{ij}^{rs}$ which is large for those pairs (i,j) which are included in the same group in many classifications and small in the opposite case. We may say that a_{ij} is a measure of "closeness" or "similarity" as regards (i,j). We may now rewrite (4.1) once more

$$D = N \sum_{i,j} \theta_{ij}^{*s} - 2 \sum_{i,j} \theta_{ij}^{*s} a_{ij} + \sum_{i,j} a_{ij} \qquad (4.1')$$

The last term can obviously be omitted from the optimization, since it occurs in (4.1') as a constant. Hence, the objective is to minimize

$$D = N \sum_{i,j} \theta_{ij}^{*s} - 2 \sum_{i,j} \theta_{ij}^{*s} a_{ij} \qquad (4.2)$$

by selecting $\{\theta_{ij}^{*s}\}$ in an optimal way.

Let the groups in classification $I^* = I^*(\mathrm{R,s})$ be indexed by 1,...,k,...,N so that $I^* = \{I_1,\ldots,I_k,\ldots\}$, and observe that $\theta_{ij}^{*s} = 1$ for $i,j \in I_k$ and that $\theta_{ij}^{*s} = 0$ otherwise. Then the minimization problem in (4.2) may be formulated as

$$\mathrm{Max} \sum_k \sum_{i,j \in I_k} a_{ij} - (N/2)\Sigma N_k^2 \qquad (4.3)$$

where N_k is the number of elements in $I_k \in I^*$. From formula (4.3) we can see that two forces are operating: the objective function increases when the closeness component $\Sigma\Sigma a_{ij}$ grows; the function also increases when ΣN_k^2 is reduced, i.e., when a uniform distribution is approached. N/2 defines the threshold of trade-off between the two conflicting goals. Having observed this we can replace the second term of (4.3) with a shift parameter $0 \leq a \leq N$ [1] which expresses a selected threshold value. We can replace the optimization problem in (4.3) with the maximization of

$$U = \sum_k \sum_{i,j \in I_k} (a_{ij} - a) \tag{4.4}$$

If $a = 0$ U reaches its maximum when all elements are placed in one group, and conversely if $a = N$ the maximum obtains when every group contains just one element. Evidently, a can be used as shift or search parameter which is adjusted to give us the desired number of groups. A convenient start value may be $a = N/2$.

The algorithm consecutively unites individual elements or groups into gradually larger groups in such a way that the value of U increases for each change. This process is stopped when U does not show any further growth in value. The convergence of the process is guaranteed by the fact that U is bounded from above and at the same time monotonously non-decreasing.

4.2 An Algorithm for Finding the Permutation Matrix S

In (3.10) and (3.12) a permutation matrix is introduced. This transforms the initial trade pattern or disaggregated regional composition into an aggregate one. In (3.8) the solution was described as the outcome of the minimization of

$$I(Q,Z) = \Sigma\, q^{rs} \log q^{rs} - \sum_{I,J} \bar{z}_{IJ} \log [\bar{z}_{IJ}/d(I)d(J)]$$

The first term on the right hand side is not affected by the

[1] N is the number of initial classifications.

selection of Z, and can therefore be deleted. Hence, the problem is reduced to the maximization of the second term. Recalling that d(I) and d(J) are elements in the diagonal matrix S^TS we can formulate this reduced problem as maximization of

$$\Sigma S_I^T Q S_J \log [S^T Q S_J / (S^T S)_I (S^T S)_J] \qquad (4.5)$$

where S_I and S_J are the I'th and J'th column vectors of the matrix S, respectively. In (4.5) the objective is entirely expressed in the sought matrix S and the known matrix Q. By specifying the desired number of regions, i.e., the dimensions of the matrix $\bar{Z}$, the function in (4.5) can be solved with some heuristic search procedure (see also Roy, Batten & Lesse, 1982).

REFERENCES

Andersson, Å.E., H. Persson (1982) Modeling International Trade Flows and Specialization, *University of Umeå, Swedish College of Forestry,* WP 1983:1

Batten, D. (1982) *Spatial Analysis of Interacting Economics,* Boston: Kluwer-Nijhoff.

Batten, D., B. Johansson and M. Kallio (1983) The Analysis of World Trade in Forest Products: Part 1 - Conceptual and Empirical Issues. WP-83-50. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Britkov, V., A. Sarkisov (1983) Elements of the Data Modification System for the Modeling Data Bank. Preprint, Institute for Systems Studies, Moscow, USSR (in Russian).

Chamberlin, E.H. (1933) *The Theory of Monopolistic Competition,* Harvard University Press.

Eriksson, J. (1981) *Algorithms for Entropy and Mathematical Programming,* PhD dissertation, Department of Mathematics, University of Linköping.

Gukova, T., A. Sarkisov (1980) Regionalization Problems in Global Development Models. In collection of papers "Methods for Studying Complex Systems (Part II)" (Proceedings of the Third Conference of Young Scientists), Institute for Systems Studies, Moscow, USSR (in Russian).

Jaynes, E.T. (1957) Information Theory and Statistical Mechanics. *Physical Review,* Vol. 106 pp 620-630; Vol. 108 pp 171-190.

Johansson, B., D. Batten (1983) Price Adjustments and Multiregional Rigidities in the Analysis of World Trade. WP-83-00 (forthcoming) Laxenburg, Austria: International Institute for Applied Systems Analysis.

Johansson, B., B. Marksjö (1983) An Interactive System for Regional Analysis of Industrial Sectors. WP-83-54. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Lesse, P.F., M. Batty, and D.F. Batten (1983) A Theory of Model Evaluations: Spatial Invariance, Transformations and Aggregations, *Environment and Planning A*.

Robinson, J. (1933) *Economics of Imperfect Competition*, MacMillan.

Roy, J., D.F. Batten, and P.F. Lesse (1982) Minimizing Information Loss in Simple Aggregation *Environment and Planning A*, 14:973-980.

Snickars, F., J.W. Weibull (1977) A Minimum Information Principle: Theory and Practice. *Regional Science and Urban Economics* 7:137-168.