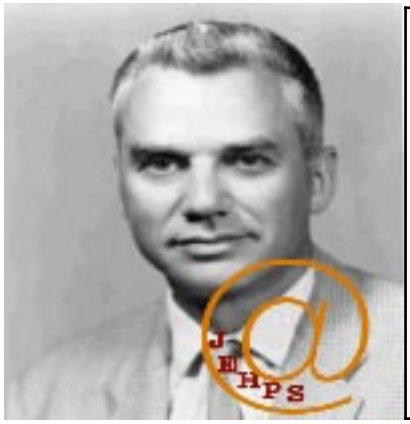

Journ@l Electronique d'Histoire des Probabilités et de la Statistique

Electronic Journ@l for History of Probability and Statistics

Vol 4, n°2; Décembre/December 2008

www.jehps.net

Some topics of current interest in clustering: Russian approaches 1960-1985

Boris MIRKIN[1] and Ilya MUCHNIK[2]

Introduction: General remarks

Early developments in data analysis in Russia, and in the Soviet Union in general have been well separated by the rather tight iron curtain through all of the covered period. However, they generally followed the international developments, because some monographs were translated into Russian, best journals reached a few libraries with a lag of one-two years, and a few Soviet scientists did attend some conferences: the international framework distinguishing between factor analysis, pattern recognition and clustering prevailed. Yet there have been a number of indigenous developments of which by far the most popular has become the structural risk minimization by V. Vapnik and Chervonenkis (Vapnik and Chervonenkis 1974) along with the concept of VC-complexity and the method of generalized portrait (Vapnik and Lerner1963) later, in 90-es, extended to the renowned Support Vector Machine (SVM) with nonlinear kernels – the latter was proposed as a tool for supervised and unsupervised learning by Bashkirov, Braverman and Muchnik (1964) under the name of "potential functions"; for further developments, mostly extending the perceptron framework, see Aizerman, Braverman, and Rozonoer (1970). Much less known are the following contributions:

- M. Bongard proposed logic predicate data analysis involving invented by him benchmark examples (Bongard 1967) that have become known as Bongard problems in cognitive psychology (see, for example, Linhares (2000));
- S. Chesnokov developed his determinacy analysis based on conditional probabilities to anticipate the analysis of associations in data mining (Chesnokov 1982);
- V. Fain proposed using techniques of the continuous group theory for describing image geometry changes (Fain 1970);

[1] School of Computer Science and Information Systems, Birkbeck University of London, UK and Division of Applied Mathematics and Informatics, Higher School of Economics, Moscow, Russia

[2] Department of Computer Science and DIMACS, Rutgers University, Piscataway, NJ, USA

- V. Finn proposed the JSM (John Stuart Mille) method involving the concept of similarity between two complex objects as the set of all subsets (substructures) in the objects' overlap (Finn 1983);
- G. Lbov proposed using logic decision functions for mixed scale data in main problems of data analysis (Lbov 1979);
- B. Mirkin proposed using the partition metric space as a tool to address main problems in categorical data analysis (Mirkin 1969) later extended to the case of mixed scale data (Mirkin 1985);
- B. Mirkin and I. Muchnik developed, in a friendly competition, an approach for approximating large networks with smaller graphs (Mirkin (1974, 1976, 1981), Muchnik (1974), Muchnik and Oslon (1980), Braverman and Muchnik (1983));
- J. Mullat and I. Muchnik proposed an original class of easily optimized order-structure "quasi-convex" set functions for finding representative layered subsets ("monotone systems" of Mullat (1976), Kuznetsov and Muchnik (1983));
- A. Orlov proposed a probabilistic framework for non-numeric data (Orlov 1979);
- L. Rastrigin and R. Ehrenstein proposed using ensembles of classification algorithms for achieving better accuracy (Rastrigin and Ehrenstein 1978);
- N. Zagoruiko and E.Vitiaev proposed a formal logic framework compatible with deriving regularities from empirical data (Zagoruiko 1979);
- Y. Zhuravlev proposed a framework of numerical characteristics of minimal sets in binary data tables, the so-called "deadlock tests" (Zhuravlev, Yunusov 1971), later extended to an algebra-based formalism (Zhuravlev 1978), and many more.

Each of the mentioned directions attracted a following and generated many papers, several dozen in some cases, mostly published in obscure typographically poor or very poor collections and conference proceedings comprising much wider subjects such as "science and technology" or "automating the national economy", though some journals did publish papers on data analysis, most notably "Automation and Remote Control" (Institute of Control Problems, Moscow) – in its division titled "Modelling intelligence and behaviour"[3]. Regular seminars led by S. Aivazian (CEMI Moscow), M. Aizerman (IPU Moscow) and N. Zagoruiko (IM Novosibirsk) helped in developing some common themes and approaches. General meetings were rather random and rare (for example, E. Braverman and I. Muchnik organized Meeting on supervised and unsupervised learning (1970, Kalinin – currently Tver) and B. Mirkin organized a couple of seminars on data analysis in sociology (1977, Cheliabinsk; 1979, Ulan-Udeh)), with two notable exceptions: A. Aivazian (Moscow) organized bi-annual conferences on "Applied Statistics" held alternatingly in Estonia (Tartu, local organiser M.-A.Tiits) and Armenia (Tsakhkadzor, local organiser V. Mkhitarian), starting from 1977, and N. Zagoruiko (Novosibirsk) organized tri-annual meetings on "Machine Discovery of Regularities" held in different places starting from 1976.

Authors of this paper already have had an opportunity of reviewing the Russian developments in data analysis (Mirkin and Muchnik 1996) – there is no need to replicate the review here. Instead, this paper concentrates on several topics of current interest in clustering and highlights the relevant Russian developments in the period 1960-1985. The topics reflect the authors' research interests and are as follows:

- Comparing clusterings
- Consensus clustering
- One cluster clustering
- Biclustering
- Network structuring

[3] Two relevant Russian journals have had English versions: (i) "Cybernetics" (Kiev) translated as "Cybernetics and Systems Analysis" (Plenum, New York), and (ii) "Computational Mathematics and Mathematical Physics" (Moscow).

1. Comparing clusterings

This subject received much attention in Russia in that period. Motivated by the axiomatic introduction of a distance measure in the set of all rankings on a finite set by Kemeny and Snell (1962), Mirkin and Cherny (1970) proposed similar axioms to introduce a distance between partitions as the Hamming distance between matrices of their equivalence relations. Consider a set I comprising N elements and a partition $S=\{S_1, S_2, ..., S_K\}$ consisting of K non-overlapping classes S_k that cover all of I. Let us define an $N\times N$ binary matrices $s=(s_{ij})$ for the equivalence relation of "being in the same class of S", by $s_{ij}=1$ if $(i,j)\in S$, and $s_{ij}=0$, otherwise. The distance between partitions R and S is defined then as the squared Euclidean or just city-block[4] distance between the matrices, $d(R,S)=\Sigma_{i,j\in I}|r_{ij}-s_{ij}|$. This distance is expressed through R and S cross-classification by Mirkin and Cherny (1970) in a currently well known way, which is the complement to the popular Rand coeffiient (Rand 1971) to unity. We are going to elaborate on this by citing a result from Mirkin (1976) that was not described in English before. There are two characteristic properties:

(i) *Independence* of the distance on coinciding parts of R and S, and

(ii) *Additive* property, $d(S,T)=d(S,R)+d(R,T)$, with respect to each partition R being between S and T, that is, satisfying inequalities $min(s_{ij}, t_{ij}) \le r_{ij} \le max(s_{ij}, t_{ij})$ for all $i, j\in I$.

It is proven in Mirkin (1976, p. 45-48) that these two define a broad class of distance measures if supplemented by a calibrating property demanding that the distance between two trivial partitions, O consisting of N singletons and U consisting of one class with all elements, $U=\{I\}$, can be expressed through a monotone increasing function $\varphi(N)$ on integers, such that $\varphi(N)>N\varphi(1)$ for all N, as follows:

(iii) *Calibrating*: $d(O,U)=\varphi(N)-N\varphi(1)$

Specifically, a symmetric function $d(S,R)$ satisfies properties (i), (ii) and (iii) if and only if

$$d(S,R)=\sum_{k=1}^{K}\varphi(N_{k.})+\sum_{l=1}^{L}\varphi(N_{.l})-2\sum_{k=1}^{K}\sum_{l=1}^{L}\varphi(N_{kl}) \qquad (1)$$

The notation in (1) refers to $S=\{S_1, S_2, ..., S_K\}$, $R=\{R_1, R_2, ..., R_L\}$ with the number of elements in S_k, R_1 and $S_k \cap R_1$ denoted by $N_{.k}$, $N_{l.}$ and N_{kl}, respectively ($k=1,..., K, l=1,..,L$).

With $\varphi(N)=N^2$, $d(S,R)$ in (1) is the double Mirkin-Cherny distance, and with $\varphi(N)=Nlog(N)$, $d(R,S)$ in (1) is the information distance recently derived by Meila (2007).

Even more impressive measure – of the degree of independence between partitions that are possibly incomplete, was derived by Plotkin (1980); this is a cubic function of the contingency table entries; unfortunately no further analysis followed.

Rauschenbach (1982) characterized the currently popular Jackard distance between subsets. For any non-empty $S, T\subset I$, Jackard index is defined as $cj=|S\cap T|/|S\cup T|$, the number of elements in the overlap of S and T related to the number of elements in their union, and Jackard distance is $dj=1-cj$. It appears, the following three properties of a measure $d(S,T)$ held for all nonempty subsets S and T:

(a) if S is part of T, then $d(S,T)=c_T|T-S|$ where c_T is a constant depending on T only;

(b) $d(S,T)\le 1$; and if S and T do not overlap, then $d(S,T)=1$;

(c) $d(S,T)=d(S,S\cup T)+d(S\cup T,T)$;

make d to coincide with dj; this was extended to the case of fuzzy subsets as well (Rauschenbach, 1982).

[4] These are the same in this case, which took several years to get noticed.

2. Consensus clustering

Consensus clustering is an activity of summarising a set of clusterings into a single clustering that represents all of them in a best way, which is utilized currently in the analysis of gene expression data (see, for example, Swift et al. 2004, Xiao and Pan 2007). This problem, for the case of clusterings being partitions was first analyzed by B. Mirkin both by means of axiomatic analysis (see a review in Day and McMorris 2003) and approximation analysis (Mirkin 1974). In the latter the problem was considered as follows: given n partitions $S_1,\ldots, S_n$ on I, find such a partition R that minimizes $\Sigma_t d(R,S_t)$ where summation goes over all t from 1 to n. It appears, this problem can be reformulated in terms of the consensus matrix M of similarities between elements of I: denote m_{ij} the number of partitions S_t in which elements $i,j\in I$ belong to the same class and take threshold $\lambda=n/2$. Then the optimal partition R maximizes the summary within cluster similarities short of λ (Mirkin 1974)

$$f(R, \lambda) = \Sigma_k \Sigma_{ij\in Rk}(m_{ij}- \lambda) = \Sigma_k \Sigma_{ij\in Rk} m_{ij} - \lambda \Sigma_k |S_k|^2 \qquad (2)$$

The threshold makes it desirable to put i and j together if $m_{ij}>\lambda$, to increase $f(R, \lambda)$, and separately if $m_{ij}<\lambda$, though this may be affected by contradicting preferences on related elements. Criterion (2) has been utilized, with $\lambda=n/2$ and no references, in the literature, see, for example, Swift et al. (2004).

Back then the current authors analyzed a test benchmark example that criterion (2) fails (Mirkin 1976). The test assumes a proper clustering, $R=\{R_k\}$ of K classes, $k=1,2,\ldots, K$, to be pre-specified on I. Now take $n=K$ and define S_k as a two cluster partition consisting of two clusters, R_k and I-R_k ($k=1,\ldots, K$). Intuitively, it is the pre-specified R that should maximize (2), which is true indeed, but only at $K \le 4$; at $K>4$ the optimal solution is always the universal U consisting of the only cluster I of all elements (Mirkin 1976). Therefore, the concept was extended to solve the issue, by assuming that each of the partitions S_t has a positive weight, w_t, so that the consensus partition R must be accompanied with a weight, w, too, either pre-specified or to be found by minimizing the accordingly updated criterion $\Sigma_t d(wR,w_tS_t)$ in which distance d is extended to be the sum of squared differences between the weighted equivalence matrices corresponding to the partitions (Kupershtokh et al. 1976). Either pre-specified $w=2(K-1)/K$ or optimal w in the benchmark example will solve the issue so that the original R is (weighted) consensus for any K.

This approximation approach was further extended by Mirkin and Muchnik (1981) with a more conventional representation of partitions using the corresponding incidence matrices rather than the square equivalence relation matrices. Given a partition $S=\{S_1, S_2, \ldots, S_K\}$, the binary $N\times K$ incidence matrix $X=(x_{ik})$ where $x_{ik}=1$ if $i\in S_k$ and $x_{ik}=0$, otherwise, defines, first, a linear space, its span $L(X)$ as the set of vectors Xa for all K-dimensional as, and, second, the orthogonal projection $P_X = X(X^TX)^{-1}X$ onto $L(X)$. Matrix P_X is of dimension $N\times N$ and can be considered a similarity matrix on set I; moreover, its elements $p_{ij}=0$ if i and j belong to different classes of S, and $p_{ij} =1/|S_k|$ if $i,j\in S_k$. Mirkin and Muchnik (1981) introduce two consensus partition concepts. Given n partitions $S_1,\ldots, S_n$ on I, with their incidence matrices $X_1,\ldots, X_n$, find such a partition R, with its incidence matrix Z, that minimizes

$$\Sigma_t ||X_t - P_Z X_t||^2 \qquad \text{or} \qquad \Sigma_t ||Z - P_{X_t} Z||^2$$

over all possible partition incidence matrices Z, where $||D||^2$ is the squared conventional Euclidean norm, that is, the sum of all D entries squared. The first criterion leads to what can be referred to as the *source consensus partition* whereas the second corresponds to the *target consensus*.

Mirkin and Muchnik (1981) reformulate these criteria in terms of both entities and features. It appears the former criterion is equivalent to the conventional clustering square-error criterion in the space of dummy

variables represented by columns of matrices X_t, whereas its reformulation in terms of similarities leads to another conventional criterion: maximize the total weighted within cluster similarities

$$g(R) = \Sigma_k(\Sigma_{ij \in Rk} b_{ij})/|R_k| \quad (3)$$

where $R=\{R_1, R_2, ..., R_K\}$ is the sought consensus partition and b_{ij} is the sum of (i,j)-th elements of the projection matrices onto all spaces $L(X_k)$, $k=1, ..., K$.

Criterion (3) was fitting into a tradition:

(i) this criterion, for arbitrary similarities, had been experimentally chosen from a number of candidate criteria by Braverman et al. (1971), and

(ii) the similarity b_{ij}, defined as the sum of inverse frequencies of the features on which i and j coincide, had been advocated as early as in 1930s by entomologist E.S. Smirnov (1898-1972), Head of the Entomology Department at the Moscow State University from 1940-72, one of the early enthusiasts of the numerical taxonomy, of which he managed at last to publish a monograph (Smirnov 1969).

The target consensus criterion reformulated in terms of similarities has the format of criterion (2), though with different similarities and threshold (Mirkin and Muchnik 1981).

It should be noted that there are two different frameworks for consensus partitioning: one is of partitions S_t produced by different versions of the same (or different) clustering algorithms on the same data, the other of categorical features t such that classes of partitions S_t correspond to their different categories ($t=1,2,...,n$). Intuitively, one may think that criterion (2) suits the former whereas criterion (3) is good for the latter. Yet no explicitly stated theoretical framework has been proposed, and the two criteria hang on anticipating the demarcation.

3. One cluster clustering

The currently popular idea that the entity set may not necessarily be meaningfully partitioned into clusters but rather just one or two clusters would suffice, while leaving the other entities unassigned, was very prominent from the very beginning of cluster analysis, manifesting itself, for example, in the concept of B-cluster (Harman and Holzinger, 1941). It was prominent in the Russian research, too. We are going to present five different approaches: defined cluster (Apresian, 1966), moving cluster (Elkina and Zagoruiko 1966), approximate cluster (Mirkin, 1976), layered cluster (monotone system) (Mullat 1976, Kuznetsov, Muchnik, 1983), and logic taxon (Lbov and Pestunova, 1985).

3.1. Apresian's cluster

Given a dissimilarity matrix $(d(i,j))$, $i,j \in I$, a subset S is referred to as an A-cluster if for any three different $i,j,k \in I$ such that $i,j \in S$ and $k \notin S$, $d(i,j) \leq d(i,k)$ (Apresian 1966). As is currently well-known, the set of A-clusters forms a hierarchy, that is, if two A-clusters overlap, then one must be part of the other. For the time being, this has been well extended by changing the cluster definition, of which probably the most popular is the concept of weak cluster in which $d(i,j) \leq max(d(i,k), d(j,k))$ (Bandelt and Dress 1989).

3.2. Forel: moving cluster

Given a pre-processed data matrix $Y=(y_{iv})$, where $i \in I$ are entities and $v \in V$ are variables, pre-specify a "cluster radius distance" $D>0$ and define tentative cluster S as the set of entities that are closer to the data grand mean than D, that is, $S=\{i : d(i,g)<D\}$, where $d(i,g)$ is Euclidean distance between row $i \in I$ and grand mean g. Now iterate the following: calculate the center of gravity in S, gs, and redefine S to be "around" the new center, $S=\{i : d(i,gs)<D\}$ – until convergence. If needed, one may find more clusters by repeating the procedure after removing the found cluster(s). The radius D can be chosen as "most suitable" of several trials. This algorithm is referred to as Forel, an abbreviation of the title, "Formal element" (Elkina and Zagoruiko 1966). Forel leaves many entities out of a few big clusters which was considered back then a weakness, but nowadays, with the one cluster perspective having received good footing, this seems more like an advantage.

3.3. Approximate cluster

Given an entity-to-entity similarity matrix $B=(b_{ij})$, specify cluster S to be found with two items, the binary membership vector $s=(s_i)$ in which $s_i=1$ if $i \in S$, and $s_i=0$, otherwise, and intensity weight λ, typically positive, and find it in such a way that the total squared difference $L(S)= \Sigma i,j \in I (b_{ij} - \lambda s_i s_j)^2$ is minimized (Mirkin 1976, Mirkin 1985). A local optimization method from these publications, in two versions related to the case of pre-specified and optimized λ, ADDI and ADDI-S, respectively, has been described in English by Mirkin (1996). Its advantage is a proven tightness of the resulting S: for every $i \in S$, its average similarity to S is greater than $\lambda/2$, and for every $i \notin S$, its average similarity to S is less than $\lambda/2$ (Mirkin 1976). A part of ADDI, related to the use of soft threshold $\pi=\lambda/2$, was proposed (with no references to the previous work) by Ben-Dor et al. (1999) as algorithm CAST which became popular in bioinformatics.

3.4. Layered cluster (monotone system)

A rather original approach, using element-to-subset linkages rather than just element-to-element similarities, was proposed by Mullat (1976) and further developed by Muchnik and his collaborators (Kuznetsov, Muchnik, 1983, Aaremaa 1985 and many more). Consider a linkage function $f(i,S)$ between all elements $i \in I$ and subsets $S \subset I$ as the input data. Typically, such a function can be meaningfully defined over any type of raw data, be to digital image or text. For example, given a non-negative similarity (weighted graph) matrix $A=(a_{ij})$, one could define $f(i,S)$ as $\Sigma_{i \in S} a_{ij}$ or $\min_{i \in S} a_{ij}$ or in many other ways so that f is monotone over S, that is, (i) $f(i,S) \leq f(i,S \cup T)$ for all S and T, in the former case, or (ii) $f(i,S) \geq f(i,S \cup T)$ for all S and T, in the latter case. A monotone linkage function $f(i,S)$, in the case (i) for certainty, leads to a set function $F(S)=min_{i \in S} f(i,S)$ characterizing the "weakest link" in S. Any weakest link function $F(S)$ can be maximized in a greedy-wise manner, leading to a string of greedy choices of individual entities, which not only specifies the maximally dense set S as a suffix of the string but also its less dense "shells" being longer suffixes. The weakest link functions are, in fact, those and only those that satisfy the condition of quasi-convexity, $F(S \cup T) \geq min(F(S),F(T))$, for all $S, T \subset I$. An indicator of a subset $T \subset I$, function G defined by the rule that $G(S)=0$ for all S that differ from T and $G(T)=1$, satisfies this condition too, but the problem of its maximization is NP-complete, which seems at odds with the above. In fact, it is not – because the corresponding linkage function $g(i, S)$ provides a great deal of information of G, which is not conventionally available. There is a deep underlying order structure in the weakest link functions (Mirkin and Muchnik 2002) and the concept fits well into specifics of organizational control (Kuznetsov, Muchnik, 1983) – this is why we think that this approach has potential for further development.

3.5. Logic taxon.

Any predicate comprised of features, say "$y_1=3$ and $y_2/y_3>5$", corresponds to the subset S of entities satisfying it. (Note: no restriction on feature scales is imposed here!) The proportion of S in I is the observed frequency f. On the other hand, predicates related to the individual statements, "$y_1=3$" and "$y_2/y_3>5$" in our example, have their frequencies on the data too, say $f1$ and $f2$. From these individual frequencies, one can easily derive the expected frequency of the combined predicate ef, according to the probability rules for Boolean operations (in our example, ef is just the product $f1*f2$). The greater the difference $f - ef$, the better S is. This criterion, as well as a heuristic for optimizing it, was proposed by Lbov and Pestunova (1985) reflecting earlier work by G. Lbov. A similar, and somewhat more straightforward criterion, maximizing the ratio $P(y_1, y_2, ..., y_n)/(P(y_1)P(y_2)...P(y_n))$, was later utilised by the founders of Megaputer Intelligence, one of a very few successful science intensive international companies launched by Russians after the collapse of the USSR (Kiselev et al 1999).

4. Biclustering

4.1. Bicluster

Biclustering is a concept related to methods that cluster rows and columns of a data table simultaneously (the term coined by Mirkin 1996). This currently is the domain given almost exclusively to single biclusters (see, for example, for a recent review Prelic et al. (2006)). Mirkin and Rostovtsev (1978) define a bicluster over a similarity table $B=(b_{ij})$ for which set of rows, $I=\{i\}$, and set of columns, $J=\{j\}$, are considered different, as a pair of row set $V \subset I$ and column set $W \subset J$ characterized by an intensity matrix $A=(a_{ij})$ such that $a_{ij} = \alpha$ for $(i,j) \in V \times W$ and $a_{ij} = \beta$, otherwise. A pair (V,W) is referred to as *associated* if the summary squared difference between B and A is minimum. Rostovtsev (1982) proposed a series of normalized and non-normalized criteria for determining the number of biclusters to be found in the same similarity matrix – all based on relating the values of the criterion optimized to those found at randomly generated data and/or starting points.

4.2. Bipartition and block structure

Given a data matrix $Y=(y_{iv})$, $i \in I$ and $v \in V$, a bipartition is formed by two partitions, $S=\{S_1, S_2, ..., S_K\}$ on I and $T=\{T_1, T_2, ..., T_L\}$on V, in such a way that each block (S_k, T_l) is a bicluster ($k=1,..., K$ *and* $l=1,..., L$). This structure is frequently considered at binary or contingency data (see, for example, Nadif, G. Govaert 2005). Braverman and Muchnik have come up with a somewhat amended cluster structure, which is suitable for entity-to-feature data type, in which features (elements of V) are partitioned according to $T=\{T_1, T_2, ..., T_L\}$on V, but partitioning of rows is done not once, but L times, for each of the classes T independently, so that such a block structure is represented by partition T on V and a set of partitions $S^l=\{S^l_1, S^l_2, ..., S^l_K\}$ on I, each S^l within a "strip" corresponding to feature subset T_l ($l=1,..., L$). The originality of the approach is in that criterion for finding such a block structure involves not just entries of data matrix Y but rather the first principal components of the feature subsets T_l - this allows for much greater flexibility, thus interpretability, of the principal component analysis because the L principal components here need not be mutually orthogonal (see, for instance, Braverman et al. 1974, Braverman and Muchnik 1983). One of the most practical of Braverman and Muchnik's methods for finding the block-structure is a two stage procedure. On the first stage, a partition T of the feature set V is sought to minimize the squared within-group differences between the first principal component of feature set T_l and each of the features in T_l. On the second stage, a K-cluster partition of the entity set is sought for each $l=1,..., L$ over a single variable, the first principal component found at the first stage. As an instructive application of the method, let us describe a block structure found on a set of 85 countries over 30 features. The features were partitioned in L=2 clusters, one corresponding to the income per capita and related features, the other – to the extent of capitalization of the economy, and the overlap of two found clusterings of the set of countries was consistent with the optimal dynamics in a two-sector model of economy: to reach the maximum consumption over a period requires a policy of overwhelming investment in the beginning and, only after that, a switch to the consumption (Braverman et al. 1974). The authors also considered a least modules analogue to the method, which led to interesting extensions of the centroid method in factor analysis. Rostovtsev (1982) extended his biclustering research to a similar block structure analysis of a similarity matrix, at which he successfully utilized the criterion of maximum difference between results found on the real and random data to automate the choice of both L and K.

5. Network structuring methods

The problem of aggregation of a weighted graph into a small graph to reflect the network flow rather than just its separate or core parts should be of increasing interest currently as giant networks, such as those in the world wide web, are emerging. Meanwhile, the early interest in graph theoretic representations of social

networks (Wasserman and Faust 1994), with the current transition to the analysis of co-citation networks seems fading while the use of factor analysis and multidimensional scaling as visualization tools is on the rise (see, for example, Gest et al. 2007). Yet the power of visualization and generalization provided by graph structures is difficult to match. This is why we think that a presentation of algorithmic developments in this field can be of interest both in the historic perspective and for further work. For a given weighted graph, or the similarity or interaction $N \times N$ matrix B between graph vertices, Mirkin (1974) and Muchnik (1974) proposed methods to aggregate B into a smaller graph (S,τ) where S is set of $K<<N$ vertices, each corresponding to a subset $S_k \subset I$ of the original vertices $(k=1,..., K)$, and τ set of arcs representing the arcs of the original graph. Muchnik (1974) assumed that the structure τ is pre-specified but subsets S_k may overlap. In contrast, Mirkin (1974) assumed the structure τ unknown, but subsets S_k not overlapping. Denoting by $s=(s_{ik})$ the matrix of incidence of sets S_k, and $t=(t_{kl})$ the matrix of aggregate structure τ, the goodness-of-fit function of the structure was defined by both as the square difference between the original matrix B and the structure recovered matrix sts^T. In the follow-up work, Muchnik and his collaborators further relaxed the output structures, allowing S be fuzzy or even probabilistic and structure τ weighted (for a review, see Braverman and Muchnik 1983). Mirkin and his collaborators extended the original model to be able to determine its parameters automatically. First, Kupershtokh and Trofimov (1975) proved that the optimal graph structure τ for a given partition S is defined by positive $A_{kl}=\Sigma_{i \in Sk}\Sigma_{j \in Sl}(bij - \lambda)$, $k=1,..., K$; $l=1,..., L$; where λ is a threshold akin to that in (2), so that there is an equivalent optimality criterion, maximization of $\Sigma_k\Sigma_l|A_{kl}|$, that does not depend on τ, which brings forward all the conventional clustering schemes such as agglomeration or exchange to be utilized if adjusted to the criterion. Further relaxation of the need to pre-specify K was achieved by Mirkin (1981) with a proposal that both parts, S and τ, should independently contribute to the data recovery criterion, which leads to two independent thresholds subtracted from B and no monotone dependence of the criterion on K.

6. Conclusion

In our previous review (Mirkin, Muchnik 1996) we made a claim of rather bleak international perspectives of the Soviet developments in data analysis: “The Soviet influence has been negligible; researchers were not included in the international community, … - Soviet science has become as inconvertible as Soviet currency, and for similar reasons.” We still think so. However, having put some effort during past decade to revitalise some of those developments, we can see that – that way or the other – worthy ideas do find their way. We hope that this paper can contribute to the process of revitalising, and it might attract attention of both historians of the science and active researchers in the field.

References

R. Aaremaa (1985) On simultaneous clustering of objects and variables with monotone systems theory. In A.-M Tiits (Ed.) Theoretical and applied mathematical problems, Tartu University Press, Tartu.

S.A. Aivazian, Z.I. Bezhaeva, O.V. Staroverov (1974) Classification of Multidimensional Observations, Statistika Publishers, Moscow (in Russian).

M.A. Aizerman, E.N. Braverman, L.I. Rozonoer (1970) Method of Potential Functions in Machine Learning, Nauka Publishers, Moscow (in Russian).

Y.D. Apresian (1966) An algorithm for finding clusters by a distance matrix, Computer. Translation and Applied Linguistics, 9:72–79 (in Russian).

H.-J. Bandelt, A.W.M. Dress (1989) Weak hierarchies associated with similarity measures: an additive clustering technique, The Bulletin of Mathematical Biology 51, 133-166.

O.A. Bashkirov, E.M. Bravermann, I.B. Muchnik (1964) Potential function algorithms for pattern recognition learning machines, Automation and Remote Control, 25, 629-631.

A. Ben-Dor, R. Shamir, Z. Yakhini (1999) Clustering gene expression patterns, Journal of Computational Biology, 6, 281-297.

M.M. Bongard (1967) Pattern Recognition, Nauka Publishers, Moscow (in Russian). English translation: Spartan Books, Rochelle Park, NJ, 1970.

E. M. Braverman, A.A. Dorofeyuk, V.Y. Lumelsky, I.B. Muchnik (1971) Diagonalization of similarity matrices and finding hidden factors, in M.A. Aizerman (Ed.) Challenges for extending capabilities of automata, Institute of Control Problems Press, Moscow, 1, 42-79 (in Russian).

E. M. Braverman, N.E. Kiseleva, I.B. Muchnik, S.G. Novikov (1974) Linguistic approach to analyzing large data sets, Automation and Remote Control, 35 (11, Part 1), 1768-1788.

E.M. Braverman, I.B. Muchnik (1983) Structural Methods for the Analysis of Empirical Data, Nauka Publishers, Moscow (in Russian).

S.V. Chesnokov (1982) Determinacy Analysis of Socio-Economic Data, Nauka, Moscow (in Russian).

W.H.E. Day and F.R. McMorris (2003) Axiomatic Consensus Theory in GroupChoice and Bioinformatics, SIAM, Philadelphia, 155 p.

V.N. Elkina, N.G. Zagoruiko (1966) An alphabet for the recognition of objects, Computing Systems, 12, Novosibirsk, Institute of Mathematics Press (in Russian).

V.S. Fain (1970) Image Recognition: Continuous Group Framework and Applications, Nauka Publishers, Moscow, 299 p. (in Russian).

V. K. Finn (1983) On computer-oriented formalization of plausible reasoning in F. Bacon-J.S. Mill style, Semiotika i Informatika, 1983, 20, 35-101 (in Russian).

S. D. Gest, J. Moody, K.L. Rulison (2007) Density or Distinction? The Roles of Data Structure and Group Detection Methods in Describing Adolescent Peer Groups, Journal of Social Structure, 8(1).

K.J. Holzinger and H.H. Harman (1941) Factor Analysis, University of Chicago Press, Chicago.

L. Hubert, P. Arabie (1985) Comparing partitions, Journal of Classification, 2, 193-218.

A. G. Ivakhnenko (1969) Self Learning Systems for Recognition and Automatic Control, Technika, Kiev, 392 p. (in Russian).

J.G. Kemeny, J.L. Snell (1962) Mathematical Models in the Social Sciences, Ginn and Company, New York.

M.V. Kiselev, S.M. Ananyan, S.B.Arseniev (1999) LA - A Clustering algorithm with an automated selection of attributes, which is invariant to functional transformations of coordinates, in J.M. Zytkow, J. Rauch (Eds.) Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD 1999, Prague, 366-371.

V. Kupershtokh, B. Mirkin and V. Trofimov (1976) The sum of within-cluster similarities as a clustering criterion, Automation and Remote Control, 3, 133-141.

V. Kupershtokh, V. Trofimov (1975) An algorithm for the analysis of macrostructure of a complex system, Automation and Remote Control, 36(11), 80-90.

E. N. Kuznetsov, I.B. Muchnik (1983) Analysis of management and control functions in organizations with monotone systems, Automation and Remote Control, 44 (10, part 2), 1325-1332.

Linhares, A. (2000) A glimpse at the metaphysics of Bongard problems, Artificial Intelligence, 121 (1-2), 251–270.

G.S. Lbov, T.M. Pestunova (1985) Grouping of objects in the space of mixed scale data, in V.G. Andreenkov, A.I. Orlov, and Y.N. Tolstova (Eds.) Analysis of Nonnumeric Data in Sociology Research, Nauka Publishers, Moscow, 141-149 (in Russian).

M. Meila (2007) Comparing clusterings - an information based distance**,** Journal of Multivariate Analysis, 98 (5), 873-895.

B.G. Mirkin (1969) A new approach to data analysis in sociology, in Y.P. Voronov (Ed.) Measurement and Modeling in Sociology, Novosibirsk, Institute of Economics Press (in Russian).

B. G. Mirkin (1974) Approximation in the space of binary relations and analysis of catrgorical features, Automation and Remote Control, 9, 53-61.

B.G. Mirkin (1976) Analysis of Categorical Features, Statistika Publishers, Moscow, 166 p. (in Russian).

B. G. Mirkin (1981) An approximation criterion for the analysis of main links structure, in B. Mirkin (Ed.) Methods for Analysis of Multidimensional Economic Data, Nauka Publishers, Novosibirsk, 62-70 (in Russian).

B.G. Mirkin (1985) Groupings in Socio-Economic Research: Design and Analysis Methods, Finansy i Statistika Publisher, Moscow, 223 p (in Russian).

B. Mirkin (1996) Mathematical Classification and Clustering, Kluwer, Dordrecht, 448 p.

B.G. Mirkin, L.B. Cherny (1970) Deriving a distance measure between partitions of a finite set, Automation and Remote Control, 31(5), 91-98.

B. Mirkin, I. Muchnik (1981) Geometric interpretation of clustering criteria, in B. Mirkin (Ed.) Methods for Analysis of Multidimensional Economic Data, Nauka Publishers, Novosibirsk, 3-11 (in Russian).

B. Mirkin, I. Muchnik (1996) Clustering and multidimensional scaling in Russia (1960-1990): A review, in P.Arabie, L.Hubert, and G. De Soete (Eds.) Clustering and Classification, River Edge, NJ: World Scientific Publishing, 295-339.

B. Mirkin, I. Muchnik (2002) Induced layered clusters, hereditary mappings, and convex geometries, Applied Mathematics Letters, 15, 293-298

B. Mirkin, S. Rostovtsev (1978) Method for selection of associated groups of features, in B. Mirkin (Ed.) Models for aggregation of socio-economic data, Siberian Institute of Economics Press, Novosibirsk, 107-112 (in Russian).

I. B. Muchnik (1974) Structure analysis of proximity graphs, Automation and Remote Control, 35(9, part 2), 1432-1447.

I.B. Muchnik and A.A. Oslon (1980) Approximation of the weighted relation matrix by a structural factor, Automation and Remote Control, 41 (4, part 1), 509-515.

J.E. Mullat (1976) Extremal subsystems in monotone systems I, Automation and Remote Control, 5,130-139; II, ibid., 8, 169-178 and ibid (1977), 1, 143-152.

M. Nadif, G. Govaert (2005) Block clustering with mixture model: comparison of different approaches, in Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), http://conferences.telecom-bretagne.eu/asmda2005/IMG/pdf/ proceedings/463.pdf (visited 18.10.2008).

A.I. Orlov (1979) The Sensitivity in Socio-Economic Models, Nauka Publishers, Moscow (in Russian).

A.A. Plotkin (1980) A measure of the degree of independence between classifications, Automation and Remote Control, 41 (4, part 2), 517-523.

A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, E. Zitzler (2006) A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics 22(9): 1122-1129.

W.M. Rand (1971) Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, 66, 846-850.

L.A. Rastrigin, R.H. Ehrenstein (1978) A collective of algorithms combined for task solving, Technology Cybernetics, 1978, 2, 116-126 (in Russian).

G. V. Rauschenbach (1982) Dissimilarities in subset spaces, in B. Mirkin (Ed.) Algorithms for Data Analysis in Socio-Economic Research, Institute of Economics, Novosibirsk, 29-43 (in Russian).

P.S. Rostovtsev (1982) Determination of the complexity of the aggregate structure of the data by using statistic criteria, in B. Mirkin (Ed.) Algorithms for Data Analysis in Socio-Economic Research, Institute of Economics, Novosibirsk, 105-128 (in Russian).

E.S. Smirnov (1969) Taxonomy Analysis, Moscow State University Press, Moscow, 188 p. (in Russian).

S. Swift**,** A. Tucker**,** V. Vinciotti**,** N. Martin**,** C. Orengo**,** X. Liu and P. Kellam (2004) Consensus clustering and functional interpretaton of gene-expression data, *Genome Biology,* **5:**R94.

V.N. Vapnik and A.Y. Chervonenkis (1974) A Theory for Pattern Recognition, Nauka Publishers, Moscow (in Russian).

V.N.Vapnik and A.Y. Lerner (1963) Pattern recognition using generalized portrait method, Automation and Remote Control, 24 (6).

S. Wasserman, K. Faust (1994) Social Network Analysis: Methods and Applications, Cambridge University Press.

G. Xiao, W. Pan (2007) Consensus clustering of gene expression data and its application to gene function prediction, Journal of Computational & Graphical Statistics, 16(3),2007

N.G. Zagoruiko, Ed. (1979) Empirical prediction and pattern recognition, Computing Systems (periodic edition), 79, Mathematics Institut, Novosibirsk (in Russian).

Y.I. Zhuravlev (1978) An algebraic approach to classification and recognition problems, Cybernetics Problems (periodic), 33, 5-78 (in Russian).

Y.I. Zhuravlev, G. Yunusov (1971) A voting rule algorithm for taxonomy, Computational Mathematics and Mathematical Physics, 11 (5), 1344-1347.