Clustering and a dissimilarity measure for methadone dosage time series

CHIEN-JU LIN[1], CHRISTIAN HENNIG[1], and CHIEH-LIANG HUANG[2]

[1] Department of Statistical Science, University College London, United Kingdom
chien-ju.lin.10@ucl.ac.uk; c.hennig@ucl.ac.uk
[2] Department of Psychiatry, China Medical University Hospital, Taiwan
psyche.hcl@gmail.com

Abstract. In this work we analyze data for 314 participants of a methadone study over 180 days. Dosages in mg were converted for better interpretability to seven categories in which six categories have an ordinal scale for representing dosages and one category for missing dosages. We develop a dissimilarity measure and cluster the time series using "partitioning around medoids" (PAM). The dissimilarity measure is based on assessing the interpretative dissimilarity between categories. It quantifies the structure of the categories which is partly categorical, partly ordinal and also involves quantitative information. The principle behind the measure can be used for other applications as well, in which there is more information about the meaning of categories than just that they are "ordinal" or "categorical".

1 Introduction

Heroin is an expensive and highly addictive drug. Heroin-dependent individuals who aim at overcoming their addiction are offered a methadone maintenance therapy (MMT) for many years. The main purpose of the MMT is not to help them to achieve abstinence but to minimize the harm associated with the use of heroin. The idea of MMT is to let drug users reduce the use of heroin by addicting to methadone and then to quit the use of methadone. The effect of methadone lasts 24 hours and consequently it has to be taken on a daily basis. To date there is no clear principle for the determination of the methadone dosage. Physicians prescribe dosages based on their own intuition.

Research has been done on daily methadone dosage taken by participants. Strain et al. (1993) studied treatment retentions and illicit drugs use and found that low dose of methadone ($\leq$ 20 mg) may improve retention but were inadequate for suppressing illicit drug use. Langendam et al. (1998) observed that participants requested to stay at a lower dosage ($\leq$ 60 mg) because of fear of double addiction. Bellin et al. (1999) studied associations between criminal activity and methadone dosage and found drug users on a high dose ($\geq$ 60 mg) were less likely to return to jail. Murray et al. (2008) found that

methadone dosage might be a response to misery. Peles et al. (2007) reported that the major risk factors for depression were female gender and high dose (> 120 mg). Gossop et al. (2000) applied the K-Means clustering method with four groups to a one year follow-up study. Two groups showed substantial reductions in their illicit drug use and criminality. They concluded that in a certain group MMT was appropriate.

Ideally, drug users are expected to reduce the use of heroin by addicting to methadone and then to quit use of methadone. The dosages should consequently have a pattern in which they go up at the beginning of the treatment and later go down. This would indicate detoxification. Physicians think that participants with such a dosage pattern and a high attendance rate most likely will have a positive outcome. Therefore, our objective is to develop a method to divide participants into groups according to their behaviour, that is, patterns of daily methadone dosage, and then find the differences between the groups. By clustering, we can study the association between dosage patterns and demographic factors, the degrees of addictions and retention of MMT. Also, the dosage patterns provide the possibility of developing a guideline for prescribing a proper methadone dosage.

The problems of clustering the participants in our study are the fluctuations of dosages and missing dosages. First of all, some participants who abused heroin while receiving the MMT did not need the full dosages indicated on their prescriptions to accommodate their addictions. In fact, they took a combination of drug and methadone in order for their addictions to be satisfied, so it was not guaranteed that the observed methadone dosages represented detoxication. Secondly, missing dosages were not missing at random. They were recorded as zeroes but zero is not normally a proper description of their state of addiction. We take account of these issues and propose to categorize dosages for alleviating the fluctuations of observed dosages and for keeping the sequences of missing dosages. Also, we propose a new dissimilarity ("p-dissimilarity") that quantifies the structure of the categories and involves quantitative information. The dosage patterns will then be represented by sequences of categories.

2 Data

Plain dosage data

Daily dosages in mg for 314 participants who received MMT between 01 January 2007 and 31 December 2008 were collected. These participants were selected from a larger study using the criterion that they had not left the study before the completion of 180 days and that they had at least 70% nonzero records of taking methadone. One month is often regarded as the minimum length of receiving methadone treatment. Participants who stay in

MMT for six months are considered to be candidates who can achieve abstinence. We considered data over 180 days. Normally, participants got weekly prescriptions. They would occasionally have multiple prescriptions but only one record of dosage taken on a single day. Besides, there was a chance that participants abused drugs, so their demand for daily methadone differed. Participants were allowed to take a dosage that was lower than the prescribed dosage to avoid overdosing. Their behavior of abusing drugs was reflected by fluctuations in their dosage taken records. By and large, following a weekly prescription, a participant took methadone daily for a period of seven days. Many participants dropped out and later returned to the treatment, or missed their treatment on a number of days. This resulted in many "zero dosage" records. However, their addiction to drugs was not zero, and these records were therefore treated as "missing values".

Category-ordered data

There are two problems with the plain dosages data. Firstly, some degrees of dosage fluctuation are not meaningful. We take the assessment of the physician, reflected in the prescription, as being more meaningful of the participant's state of addiction than occasional drug use or mood changes that may have led them to select slightly lower dosages on specific days. However, daily dosages are still of some informative value because not for every dosage taken there is a unique prescription explaining it; sometimes participants had more than one prescription from more than one physician to choose from. In principle, if no drugs are abused, methadone taken by participants should show long sequences of stability. Secondly, missing dosages are to be taken into account. These missing dosages usually have a specific meaning, namely that the participant did not show up for receiving methadone. Technically, these events are recorded as zero, but they could have various meanings. In some occasions there was no change to the status of the participant compared to surrounding days at which methadone was taken; the participant was just unreliable or felt so well on the day that they did not believe that they needed methadone. However, in some cases, particularly if there were longer absences, missing values point to more severe problems of the participant, or a tendency to leave the study, or illicit drug use. In any case, this cannot be properly reflected by the value "zero". In this paper, we treat missing values as a specific category. Lin (2014) also carried out experiments with imputation. A sensible scheme for imputation is difficult to define, because it should depend on the non-missing values surrounding the missing days, and also on length of periods of missingness.

Therefore we constructed a new data format by categorizing daily dosage. In our study, cut points for categorizing dosages were defined by the physician. He suggested that dosage in the range of 20 mg could be considered virtually the same. This meant that the qualitative difference between two dosages in the same interval could be treated as irrelevant. This corresponds to the fact

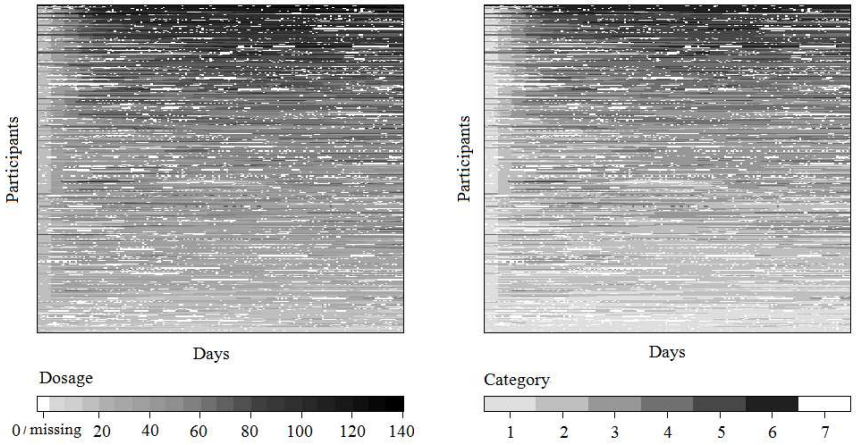

Fig. 1. Heatplot of plain dosage data and heatplot of category-ordered data
Each horizontal line represents records of a participant from day 1 to day 180. The 314 participants are ordered by the average of their dosages.

that physicians often used multiples of 20 mg in their prescriptions. We defined six categories for dosages smaller than or equal to 20 mg, 21-40 mg, 41-60 mg, 61-80 mg, 81-100 mg and greater than 100 mg, recoded as 1,. . . , 6 (Likert-coding). Another category was "missing (zero)". The resulting dataset is called "category-ordered data". This minimized the implications of irrelevant daily fluctuations and outliers (although there were no really extreme ones), and reflected the interpretation of the dosages by the physicians.

To explore the uncategorized and the categorized datasets, a heatplot is used. This is a technique to represent data by color. Each horizontal line in a heatplot represents the data of each participant. Figure 1 shows the heatplots of plain dosage and of the category-ordered data. Each horizontal line represents records of a participant from day 1 to day 180. In the graph, the 314 participants are ordered by the average of their dosages. The colour spectrums of dosage and that of category are displayed below the heatpltos. We observe that most dosage records in the first week are in category 1, as the initial prescription dosage for participants, most of which have no previous experience of the MMT, is 20 mg. Subsequently, the colours of dosage start to change, reflecting the fact that the doctors started adjusting the dosage. Also, it can be seen that most of these movements from category to category go to the next nearest category.

3 The new p-dissimilarity measure

Motivation

We define a new dissimilarity measure called "p-dissimilarity" for the category-ordered data by summing up daily dissimilarities between categories. In Lin (2014) there is a discussion of dissimilarity measures that take into account the time series structure, but as far as such measures are already in the literature, they cannot be easily adapted to our data structure (such as methods related to fitting autoregressive models to time series with continuous data), or they seem inappropriate such as "time warping" (Berndt and Clifford, 1994), because the absolute length of periods of stability is very meaningful in the context of methadone therapy whereas such lengths are treated as unimportant in time warping.

We will follow the philosophy outlined in Hennig and Hausdorf (2006), according to which a dissimilarity measure should formalise the "interpretative distance" between objects according to knowledge of the subject matter.

A specific feature of the "interpretative distance" for the data at hand is that similarity between participants is mainly governed by periods in which they are on the same dosage (category). The distinction whether categories on a day are the same or different is more important than how different they are given that they are different, because changing categories even between neighbouring categories is interpreted as indicating a substantial change in the condition of the participant. We will define a dissimilarity function that assigns a quantitative value to distances between neighbouring categories and categories further apart in a concave monotonic fashion, i.e., further categories are further away, according to the dissimilarity, but the increase of the distance becomes smaller moving further away from a category and its neighbours. This implies information that can be seen as stronger than ordinal; note, however, that any method for defining distances between ordinal categories amounts to imposing a quantitative effective distance between them. This may be governed by the distribution of the data (as when midranks are assigned or latent normality is assumed) or by the meaning of the categories and the context, which we prefer. Our dissimilarity will involve a tuning constant p, which will tune the level of information between treating the categories as purely categorical (for $p = 0$, the dissimilarity will only count how often participants are in different categories) and treating them as quantitative and equidistant ($p = 1$). The "missing value" category will be treated in a specific way, as having the same dissimilarity from all other categories.

An example for the above is that if data for three participants (A,B,C) for 7 days are [1, 1, 1, 1, 1, 2, 2], [1, 1, 4, 1, 1, 2, 2] and [1, 2, 1, 2, 1, 3, 2], we want to define a dissimilarity measure that treats A and B as more similar than A and C, so the effective distance between 1 and 4 between A and B on day 3 should not dominate the fact that A and C differ on three different days.

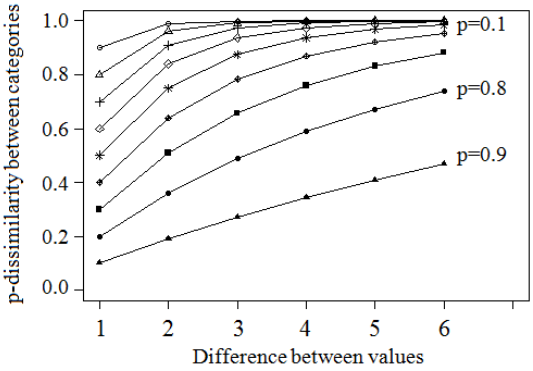

Fig. 2. p-dissimilarity between categories

Definition of the p-dissimilarity

Let $D(\cdot,\cdot)$ denote the dissimilarity between participants and $d(\cdot,\cdot)$ denote the dissimilarity between categories. Let $x_{it} \in \Theta = \{1,\ldots,\theta,\theta+1\}$ be the category-ordered data for the participant i on the t^{th} day since they joined the MMT. $\theta = 6$ is the number of Likert-coded dosage categories, and missing values are coded as $\theta + 1$. The p-dissimilarity between participants i and i' with a category for missing values is defined by

$$d(x,y) = \delta(x,y)(1-p^{\alpha(x,y)} + (1-\delta(x,y))(1-p^{\beta}), \tag{1}$$

$$D(i,i') = \sum_{t=1}^{T} d(x_{it}, x_{i't}). \tag{2}$$

where $\delta(x,y) = 1(x \le \theta, y \le \theta)$, an indicator that neither x nor y are missing, $0 < p < 1$, $\alpha(x,y) = |x-y|$, the difference between the Likert-codes of the categories, and $1 < \beta < (\theta - 1)$.

The constant p tunes the dissimilarity between categories. The p can be interpreted as a switch between data being treated as categorical and linear in the Likert codes. For $x \neq y$ and both non-missing, it can be shown that $d(x,y) = (1-p)\sum_{l=0}^{|x-y|-1} p^l$. Therefore d is monotonic in $|x-y|$ and concave, see Figure 2. For $p \approx 1$ this is almost linear in $\alpha(x_{it}, x_{i't})$, whereas for $p \approx 0$ this is close to 1 for all nonzero differences between the Likert codes.

The dissimilarity between any category and a missing value is $(1-p)^{\beta}$. Missing values can have very different meanings as explained above. Therefore missing values were not treated as particularly close to any specific category, and the constant of $(1-p)^{\beta}$ was even applied between two missing values, implying that there could be a nonzero dissimilarity between two participants with the same values on all 180 days if this included missing values. Also it means that in general the p-dissimilarity violates the triangle inequality (although this is not the case if no missing values occur). Note that Hennig

and Hausdorf (2006) argue that fulfilling the triangle inequality is not in itself a virtue of a dissimilarity measure, but only if there are subject matter reasons why it should be fulfilled for the "interpretative distance" between objects.

The parameter β tunes the dissimilarity involving missing values compared to the distances between non-missing values. $\beta = 1$ means that missing values are treated as if they were neighbouring to any category.

For practical application, the parameters p and β need to be specified. p was specified by subject matter considerations. Given the arguments before, it is clear that p should neither be very close to 1 nor very close to zero, because the very motivation for the p-dissimilarity is that a compromise between these extremes is attempted. Guided by medical considerations, we chose $p = 0.6$, for which $|x-y| = 2$ leads to a dissimilarity already very close to the maximum value, i.e. a difference of two dosage categories between participants is already implied to be very substantial, but $d(x, y)$ with $|x-y| = 2$ is still considerably larger than with $|x - y| = 1$. β was chosen as 1.42, which was the average of all $|x - y|$ occurring in the dataset between different participants on the same day with $\delta(x, y) = 1$, so missing values were treated as "in average distance to everything".

Arguments for choosing these parameters can only be imprecise. Lin (2014) carried out sensitivity analyses using other values of p and β, showing that changing p and β to values that can be seen as having a similar interpretation does not affect the clustering below much.

4 Clustering of the category-ordered data

We apply the PAM clustering method (Kaufman and Rousseuw, 1990) with five clusters and the p-dissimilarity with ($p = 0.6, \beta = 1.42$) to the category-ordered data. The clustering is then related with further information about the participants. In Lin (2014), various clustering methods (PAM, complete, average and single linkage clustering) have been compared for this dataset regarding the Average Silhouette Width (Kaufman and Rousseuw, 1990) and the prediction strength (Tibshirani and Walther, 2005), which resulted in the choice of PAM with five clusters.

Several one-way ANOVAs and χ^2-tests were run to see whether characteristics of the participants varied across clusters. There is not enough evidence to conclude that the mean participant ages differ between the five clusters (p-value is 0.084). The result of the ANOVA test about the cluster-wise mean ages of heroin onset is borderline significant (p-value is 0.062), there could be some effect of age of heroin onset. χ^2 tests show that there is not enough evidence to conclude that there exists a relationship between clusters with respect to gender ($p = 0.377$), education ($p = 0.996$), marital status ($p = 0.429$) and occupation, which is a binary variable indicating whether the participant is occupied or not ($p = 0.310$). Figure 3 shows the frequency of the categories from day 1 to day 30 for the five clusters. The y-axis indicates category and

the x-axis indicates days. The colour designates frequency. We observe the following: (1) There is an upward trend in categories over time, particular in cluster 5. (2) Cluster 1 seems to have more missing values than cluster 5. For convenience, we define the pattern of detoxification in three stages. Stage I represents that the methadone dosage goes up, stage II represents the dosage staying stable, and stage III represents the dosage going down. We attempt to summarize the pattern of detoxification for each cluster by the (rough) time point on which the three stages are observed to start: Cluster 2 (day 1-40-100), Cluster 3 (day 1-80-140), cluster 4 (1-100-150), cluster 5 (1-100-150). Also, the majority of patients in cluster 4 and cluster 5 have their dosages in high categories. This means that participants who are highly addicted to heroin might take longer to finish the detoxification process.

5 Discussion

We selected a meaningful sample with 314 participants. We took account of the weekly prescriptions, fluctuations in dosage taken records and patterns of missing dosages, and defined the category-ordered data. The final clusters were obtained by using the PAM method with five clusters and the p-dissimilarity. Unfortunately, without the data of whether participants achieve abstinence or not, we cannot understand the relationship between treatments and final outcomes. Despite the fact that none of the five clusters could very easily be distinguished in terms of, say, their demographics, the sequences of categories for the five clusters were clinically useful. We found that the heroin onset age might have an influence on the patterns of detoxification. Participants with low addiction reduced the use of heroin by addicting to methadone in the first month and attempted to reduce/quit the use of methadone by the third month. As for participants with high addictions, few attempted to reduce the use of methadone up until the fifth month and most required more time to finish the detoxification process.

Lin (2014) presents more cluster validation and discussion. She found that the data could not be significantly distinguished from a Markov model without clustering structure, which means that the observed patterns cannot be safely be assumed to correspond to a "real" clustering of methadone participants. She also computed a clustering based on raw data (instead of category-ordered data), again with PAM and five clusters. The adjusted Rand index between this clustering and the one presented here was 0.54. In any case the clustering can be used for helping physician's decision making, because they give a simple summary of the complex range of existing dosage patterns.

The p-dissimilarity is based on assessing the interpretative dissimilarity between categories and focused more on sequence of constancy and less on sudden changes in categories. This was used to measure dissimilarity between the 180-day time series of the participants. It implements concepts of variables the categories of which cannot properly be classified as purely categorical or

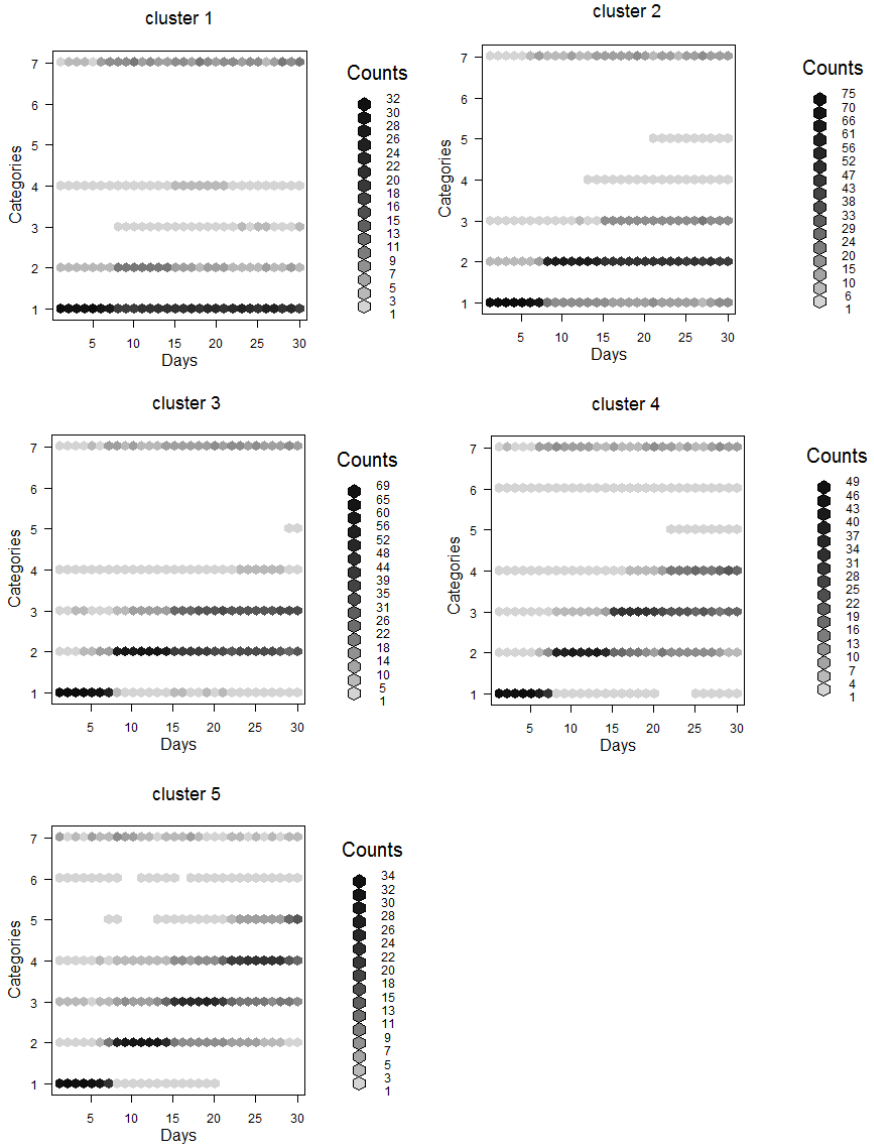

Fig. 3. Frequency of the categories from day 1 to day 30 for the five clusters. Each horizontal line represents the number of participants in each of the 7 categories from day 1 to day 30. The colour designates the number of participants.

ordinal, and can be used for incomplete data. It could be applied in wider areas of application where researchers have a quantitative idea about the interpretative distance between categories, which could be between a categorical

concept in which all differences between categories have the same distance, and a Likert-scaling concept with linearity in the Likert codes. See Hennig and Liao (2013) about related ideas for quantifying distances between categories.

References

BELLIN, E., WESSON, J., TOMASINO, V., NOLAN, J., GLI CK, A. J. and OQUENDO, S. (1999): High Dose Methadone Reduces Criminal Recidivism in Opiate Addicts. *Addiction Research, 7, 19–29*

BERNDT, D. and CLIFFORD J. (1994): Using Dynamic Time Warping to Find Patterns in Time Series. *AAAI-94 Workshop on Knowledge Discovery in Databases, 229–248*

GORDEN, A.D. (1999): *Classification*. Chapman and Hall.

GOSSOP, M., MARSDEN, J., STEWART, D. and ROLFE, A. (2000): Patterns of Improvement after Methadone Treatment: 1 Year Follow-up Results from the National Treatment Outcome Research Study. *Drug Alcohol Abuse, 60, 275–286*

HENNIG, C. and HAUSDORF, B. (2006): Design of Dissimilarity Measures: a New Dissimilarity Measure between Species Distribution Ranges. In: V. Batagelj, H-H Bock, A. Ferligoj and A. Žiberna (Eds.): *Data Science and Classification.* Springer, Berlin, 29-37.

HENNIG, C. and LIAO, T. F. (2013) How to Find an Appropriate Clustering for Mixed Type Variables with Application to Socioeconomic Stratification (with Discussion) . *Journal of the Royal Statistical Science, Series C (Applied Statistics), 62, 309–369.*

HUBERT, L. and ARABIE, P. (1985) Comparing Partitions, *Journal of Classification, 2, 193–218*

KAUFMAN, L. and ROUSSEUW, P.J.(1990): *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley and Sons.

LANGENDAM, M., VAN HAASTRECHT, H., BRUSSEL, G., VAN DEN HOEK, A., COUTINHO, R. and VAN AMEIJDEN, E. (1998): Research Report Differentiation in the Amsterdam Dispensing Circuit: Determinants of Methadone Dosage and Site of Methadone Prescription. *Addiction, 93, 61–72*

LIN, C. J. (2014) *A Pattern-Clustering Method for Longitudinal data - Heroin Users Receiving Methadone*. PhD thesis, Department of Statistical Science, University College London.

MURRAY, H., MCHUGH, R., BEHAR, E. and PRATT, E. (2008): Personality Factors Associated with Methadone Maintenance Dose. *The American Journal of Drug and Alcohol Abuse, 34, 634–641*

PELES, E., SCHREIBERA, S., NAUMOVSKYA, Y. and ADELSONA, M. (2007): Depression in methadone maintenance treatment patients: Rate and risk factors. *Affective Disorders, 99, 213–220*

STRAIN, E. C., STITZER, M. L., LIEBSON, I. A. and BIGELOW, G.E. (1993): Methadone dose and treatment outcome. *Drug and Alcohol Dependence, 33, 105–117*

TIBSHIRANI, R. and WALTHER, G. (2005): Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics, 14, 511–528*