


5-1-2017

Genetic Algorithms for Cross-Calibration of Categorical Data

Suja M. Aboukhamseen
Kuwait University, saboukhamseen@yahoo.com

Rym A. M'Hallah
Kuwait University, rymmha@yahoo.com

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Aboukhamseen, S. M. & M'Hallah, R. A. (2017). Genetic algorithms for cross-calibration of categorical data. *Journal of Modern Applied Statistical Methods*, 16(1), 722-742. doi: 10.22237/jmasm/1493599080

This Algorithms and Code is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in *Journal of Modern Applied Statistical Methods* by an authorized editor of DigitalCommons@WayneState.

Genetic Algorithms for Cross-Calibration of Categorical Data

Suja M. Aboukhamseen
Kuwait University
Kuwait City, Kuwait

Rym A. M'Hallah
Kuwait University
Kuwait City, Kuwait

The probabilistic problem of cross-calibration of two categorical variables is addressed. A probabilistic forecast of the categorical variables is obtained based on a sample of observed data. This forecast is the output of a genetic algorithm based approach, which makes no assumption on the type of relationship between the two variables and applies a scoring rule to assess the fitness of the chromosomes. It converges to a good-quality point probability forecast of the joint distribution of the two variables. The proposed approach is applied both at stationary points in time and across time. Its performance is enhanced when additional sampled data is included, and can be designed with different scoring rules or made to account for missing data.

Keywords: categorical variables, cross-calibration, genetic algorithms, probability forecasting

Introduction

Estimating the joint probability distribution of two categorical variables, based on observed data, is a common yet elusive statistical problem. Depending on the nature of the categorical variables and the intricacies that characterize their relationship, such an endeavor can be highly technical and computationally intensive. In addition, the observed data used to estimate the relationship often contains numerous sources of error or bias. Such errors, generally due to operators, equipment, or the environment, further complicate the problem; impairing the validity of any inference.

Statistical calibration models the relationship between two variables that measure the same characteristic. It saves researchers, industrials and technicians valuable time, money and effort by providing a mechanism that gives a more accurate measurement to a corrupted reading (Osborne, 1991). Its application is

Suja M. Aboukhamseen is a lecturer and researcher. Email her at saboukhamseen@yahoo.com.

particularly vital in two cases. The first case emanates when the data consists of precise measurements acquired using an invasive, destructive, costly, or time-consuming technique. In such a situation, there usually exists an alternative measurement scheme that is more complaisant but not as reliable. Paired samples from the two measurements may be calibrated; thus, providing a mechanism to forecast the more reliable method from the less reliable one. The second case arises in problems requiring data comparability. It occurs when more than one technique gives valid and reliable measurements of a certain characteristic and there is a need for cross comparison, over time or across individuals. This cross comparison or mapping or translating of one measurement of a specific phenomenon to another is known as cross-calibration.

In both cases, the data may be quantitative or qualitative (categorical). The nature of categorical data brings its own set of challenges. The data may be self-reported or may consist of self-responses/assessments. The challenge herein lies in assessing the different ways individuals apply and interpret categorical response scales (Salomon et al., 2004; Murray et al., 2002; van Buuren & Hopman-Rock, 2001). However, the calibration of such variables requires that the mapping process be customized to fit the nature of their relationship. The traits that characterize the relationship must be explicitly stated in order to maintain its integrity during the translation process. Catering to the requirements of the statistical association often means imposing restrictions on the outcomes through complex mathematical models and structures.

Assume that X and Z are categorical random variables that measure the same qualitative random phenomenon with r and c possible classes, respectively. Let π be the matrix of joint probabilities of X and Z where $\pi_{ij} = P(X = i, Z = j)$ for $i = 1, \dots, r$ and $j = 1, \dots, c$. Further assume that π is unknown, but that there exists an observed sample of N pairs of qualitative readings on (X, Z) of the single characteristic of interest. The N pairs are cross-classified into an $r \times c$ contingency table \mathbf{n} which represents the observed relationship between the categories of the two variables X and Z . In the contingency table, the cell frequency n_{ij} , $i = 1, \dots, r$, $j = 1, \dots, c$, denotes the number of readings classified simultaneously into category i by the qualitative reading on X and into category j by the qualitative reading on Z , with $\sum_{i=1}^r \sum_{j=1}^c n_{ij} = N$. Let the observed relative frequency distribution corresponding to the contingency table \mathbf{n} be denoted by \mathbf{p} where

$$p_{ij} = \frac{n_{ij}}{N}, i = 1, \dots, r \text{ and } j = 1, \dots, c.$$

GENETIC ALGORITHMS FOR CATEGORICAL DATA

The objective is to use the observed relative frequency distribution \mathbf{p} to find an estimate of the functional translation π between X and Z ; explicitly to estimate the conditional distributions $P(Z|X)$ and $P(X|Z)$. Since both distributions $P(Z|X)$ and $P(X|Z)$ are derived from the joint $P(X, Z)$, it is sufficient to find the joint probability function π . The notions behind the science of probability forecasting are used to derive an estimate of π .

DeGroot and Fienberg (1983), Dawid (1982), Schervish et al. (2014) and others established guidelines as to what constitutes a good forecasting generating system. However, how to construct that system remains an open question. In some fields, the forecasting mechanism relies heavily on expert opinion. In others, more objective procedures are employed. Herein, our focus is on the development of a forecasting generating system. A genetic algorithm (GA) -based method is applied, that searches for a (near-)optimal translation between the variables of interest. The translation corresponds to a joint distribution in the form of a probability forecasting system, from which predictive estimates of one of the two variables may be generated for a specific set of values of the other variable.

A primary advantage of this approach is that it obtains this translation without explicitly accounting for constraints that characterize the nature of the relationship between the variables. It uses the observed sample data to guide the search process. Specifically, the GA fitness construct, which is based on methods developed in probability forecasting theory (DeGroot and Fienberg, 1983; Lichtenstein et al., 1982; Gneiting and Katzfuss, 2014), ensures that the generated forecasts are valid and that they are the best among all forecasts in their class.

The purpose of this study, therefore, is to provide an overview of applications of cross calibration and genetic algorithms, and to propose a genetic algorithm. To further improve the reliability of the generated estimates, a quasi-Markov element is added to the analysis. It extends the method to cross-calibrate categorical variables measured longitudinally over time, where the calibration forecasts are generated both forward and backward on a time scale. Incorporating time broadens the applicability of the methodology. It models the relationship in a manner that is closer to the true state of nature, thus enhancing the accuracy of the estimates. This is supplemented with an illustrative example using stroke rehabilitation data.

Background

Applications of Cross-Calibration

The importance of cross-calibration emanates not only from the savings it induces, but mainly from its wide areas of applicability. Applications for this kind of analysis are manifold, making statistical calibration a valuable analytical tool. Possible fields of applications are demography, psychology, engineering, and item response theory. The following presents many fields requiring cross-calibration.

In surveys, cross-calibration facilitates the comparison of results from different questionnaires and the evaluation of response consistency. In corrosion analysis, a fundamental part of engineering, pipes and wires of oil fields are subject to corrosion because of harsh weather conditions. Following up the progress of corrosion is essential not only to production and transport of oil products but most importantly to the safety of the equipment and the personnel. Accurate tests for the state of corrosion are often invasive, destructive, and costly. The use of statistical calibration provides an efficient cost-effective alternative.

In the computation of official statistics, indicators are essential in monitoring and assessing the performance of a nation's public policy agenda, development, and how far a nation has come along in attaining its goals. Because the concepts stated above are intuitively understood, standards for their computation and compilation tend to vary widely depending on the country and the era. This makes the comparison of indicators either among countries or over time within countries exceedingly difficult. In light of today's United Nations' millennium goals, many nations are eager to show how far they have come towards attainment. This is only possible through valid data comparison, which is achievable via cross-calibration (Murray et al., 2002).

Similarly, in medicine, the assessment of a given treatment may be conducted differently depending on the researcher's preference or the time in which the study was carried out. The development of a quantitative translation between them enables the comparison of clinical trials in particular those requiring a longitudinal design over time (van Buuren et al., 2001).

In psychometrics, evaluating people's abilities, attitudes, and cognition through the process of testing and scoring is essential. Item response theory (IRT) is used in psychometrics to develop and refine tests that measure latent traits of individuals. The development of reliable techniques to measure traits such as intelligence and scholastic aptitude are of primary aim/essence of common exams, and tests of certification, such as the GRE and GMAT exams. Calibration is used

in IRT to provide a frame of reference to interpret test results, to equate tests, and to unify measurement scales both within the test items of a single test and between tests. The current practice in many of these applications is limited in scope. In some, such as IRT, the analysis requires impractical and unrealistic assumptions of independence between the items (categories) under investigation. Other applications require complicated models that tailor each aspect of the relationship separately and impose assumptions that are at many times invalid. As a result, the translations produced by the calibration model may be deficient and inaccurate. The proposed method overcomes these pitfalls by applying a methodology that makes no assumption on the type of relationship between the categorical variables under consideration.

GA Applications in Statistics

GAs mimic the role nature plays in refining and improving creation. GAs apply selective procreation and survival of the fittest to produce (near-)optimal solutions. They start from an arbitrary initial population consisting of a set of K chromosomes, where each chromosome k , $k = 1, \dots, K$, acts as a representative solution to the problem. The population undergoes an iterative process of selection, crossover, mutation, and survival of the fittest to form future generations; thus, instigating an artificial evolutionary process. The algorithm iterates until it satisfies a stopping criterion, which can be a prefixed number of iterations without improvement (i.e., convergence of the fitness function), a time limit, or a preset number of generations, n_g .

Many fields of science, such as bio-informatics, computer science, genetics, operations research, economics, engineering, quality control and mathematics, have benefited from GA's straightforward yet efficient solution strategies. GAs identified (near-)optima to numerous practical problems with varying degrees of complexity. Sayed et al. (2009) show that GAs and their hybrids can improve the predictive performance of regression models. Chen et al. (2015) apply an adaptive GA to forecast the holiday daily tourist volume based on seasonal tendency. Huang et al. (2014) used GAs to assess the quality of a certain type of salted meat based on three quality indices whose values are inferred from a colorimetric sensor array. Stojanovic et al. (2013) apply a self-adjusting GA to model the behavior of dams. Liu et al. (2013) develop a real-time GA that forecasts water quality in river crab aquaculture. Nieto et al. (2013) forecast the presence of cyanotoxins in the Trasona water reservoir of Northern Spain via GAs.

Örkcü (2013) construct a hybrid GA to choose the minimal subset of explanatory variables of a multiple linear regression model. Wibowo and Desa (2012) employ GA in conjunction with kernel principal component analysis to predict the non-linear relationship between surface roughness resulting from milling processes and the milling machine parameters in the presence of multiple collinearity. Huang (2012) designs a support vector regression GA for stock selection. Ahn et al. (2012) use GAs to forecast the appraisal value of a real estate. Aydılek and Arslan (2013) identify missing values in data sets via GAs.

In the field of scientific calibration, GAs are applied to estimate model parameters and generate predictions (Vitkovský et al., 2000). However, the application of GAs to statistical calibration in general and to categorical cross-calibration in particular remains limited.

Procedure

Although the observed relative frequency distribution \mathbf{p} is a valid statistical point estimate of π , the true joint probability function of X and Z , it may, in many instances, be biased or corrupted because it is subject to numerous sources of errors. To obtain an alternative point estimate of π based on the same observed sample frequency distribution \mathbf{p} , a GA evolutionary procedure is applied for categorical data. Unlike most GAs, the proposed GA design does not require encoding the data and maintains the data's structural integrity throughout the execution of the algorithm.

Chromosome's Definition and Fitness

When considering unknown outcomes from categorical variables, a common tacit employed is the probability of occurrence in each category. When generated for a future event, this probability is a point probability forecast. If the probability of occurrence is evaluated for each forecast category, then the sum of the probabilities should equal one; constituting a probability forecasting system. Given the available information, a probability statement about the unknown outcome of a categorical variable can be calculated and its competency evaluated.

Of the numerous criteria that are available to assess probability forecasts, (i.e. validity, refinement, etc.), calibration and scoring rules defined on the probabilities and their subsequently observed outcomes are among the more prevalent methods (Dawid, 1982; Gneiting & Katzfuss, 2014). A scoring rule is

GENETIC ALGORITHMS FOR CATEGORICAL DATA

the squared error function in which scores for all the forecast probabilities are aggregated and averaged to evaluate the system's predictive performance.

Even though originally developed for subjective probability forecasting in the field of meteorology, subjective probability forecasting has a broad applicability and a wide range of applications. For instance, it can be applied to cross-calibration and incorporated into the proposed GA as follows. For our purposes, we regard the GA chromosome k in generation g as an expression/propagation of some objective forecasting system $\hat{\pi}_k^g$. In this regard, the chromosome forecasting performance may be assessed and compared with other chromosomes.

GA, which is sequential in nature, obtains K possible estimates of π at each iteration (or generation), $g = 1, \dots, n_g$. The $r \times c$ relative frequency matrix for the two categorical variables X and Z , $\hat{\pi}_k^g$, for each chromosome $k, k = 1, \dots, K$, of iteration g is a possible estimate of π . The relative frequency $\hat{\pi}_{ijk}^g = \frac{\hat{\eta}_{ijk}^g}{N}$, represents the k^{th} probability forecast of $P(X = i, Z = j)$ at iteration g , where $\hat{\eta}_{ijk}^g$ are realizations from the k^{th} proposed joint probability $\hat{\pi}_{ijk}^g = P(X = i, Z = j)$ at iteration g of the number of times $X = i$ and $Z = j$. The sum of all frequencies, $\sum_{i=1}^r \sum_{j=1}^c \hat{\eta}_{ijk}^g$, which equals N , is independent of g and k . Thus, the sum of all relative frequencies, $\sum_{i=1}^r \sum_{j=1}^c \hat{\pi}_{ijk}^g$, always equals 1.

The fitness of a chromosome depends on the fitness of its genes. It reflects how well-calibrated the forecast frequency $\hat{\pi}_{ijk}^g$ is in comparison to p_{ij} , the observed proportion of times that $X = i$ and $Z = j$ in the observed data. A probability forecast is considered well calibrated if $\hat{\pi}_{ijk}^g = p_{ij}$. The larger the discrepancy between the observed relative frequency and the forecast probability, the less-calibrated the gene. Hence, the chromosome fitness $F_k^g, k = 1, \dots, K$, is gauged by the scoring rule

$$F_k^g = \frac{1}{N} \sum_{i=1}^r \sum_{j=1}^c (\hat{\pi}_{ijk}^g - p_{ij})^2,$$

which is the sum of the squared differences of the observed and forecast frequency. The chromosomes within the population are hitherto evaluated and ranked according to this criterion. The fitness function $F_k^g, k = 1, \dots, K$ is a proper

scoring rule (Brier, 1950). Therefore, it ensures the sharpness and calibration of the probability forecasts of the selected chromosome.

GA's Design

The proposed GA's design follows. The initial population consists of K randomly generated chromosomes. Only the fittest $\frac{K}{2}$ chromosomes of the population are granted procreation or crossover privileges. The other least fit $\frac{K}{2}$ chromosomes are deemed too weak and, therefore, unworthy of mating.

Crossover combines the genes of two existing chromosomes to generate two offspring. First, two chromosomes are selected to become parents, $Parent_1$ and $Parent_2$. Second, two integers s_1 and s_2 are randomly generated from the discrete intervals $[1, r]$ and $[1, c]$, respectively. Third, the sub-matrix consisting of the first s_1 rows and the first s_2 columns is cut out of $Parent_1$ and positioned on the same location on $Parent_2$, thus producing $Child_1$. This new offspring consists of the intersection of the first s_1 rows and s_2 columns of $Parent_1$ and of all other entries of $Parent_2$. Simultaneously, a sub-matrix of the same size and location is removed from $Parent_1$ and inserted into $Parent_2$ in the same way, giving rise to a second offspring, $Child_2$. This latter has the reverse composition of $Child_1$ with the sub-matrix of its first s_1 rows and s_2 columns emanating from $Parent_2$ and the remaining entries from $Parent_1$. Figure 1 illustrates the crossover of $Parent_1$ and $Parent_2$ to produce two children $Child_1$ and $Child_2$. The chromosomes are 5×3 matrices; i.e., categorical variables X and Z have 5 and 3 classes, respectively. The crossover chooses the two integer numbers $s_1 = 3$ and $s_2 = 2$ from the discrete uniforms $[1,5]$ and $[1,3]$, respectively. The light grey shaded areas of the parent chromosomes combine to form $Child_1$ and the dark grey shaded areas constitute $Child_2$.

To preserve the uniformity and hence the coherence of the new offspring, the alleles within $Child_1$ and $Child_2$ must be re-scaled. This requires that the relative frequencies in the child add up to 1. This is done by dividing each relative frequency by the existing total. The offspring are then merged with the existing population of generation g which consists of the $\frac{K}{2}$ parents that were involved in crossover and the $\frac{K}{2}$ childless chromosomes. The merged population has $\frac{K^2}{2}$

GENETIC ALGORITHMS FOR CATEGORICAL DATA

chromosomes: the K chromosomes of generation g and the $2 \frac{K}{2} \left(\frac{K}{2} - 1 \right)$ offspring chromosomes. The merged population is then assessed and ranked.

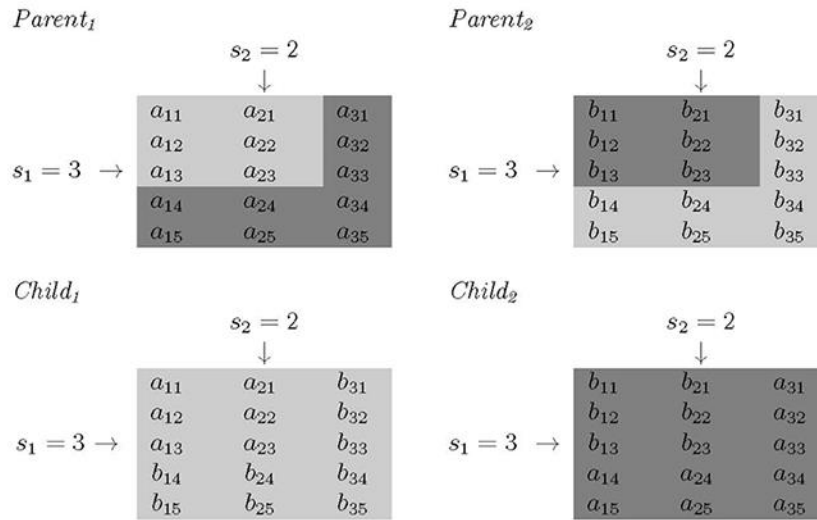


Figure 1. Crossover of two 5×3 parent chromosomes with $s_1 = 3$ and $s_2 = 2$ crossover points.

Further evolution of the population is enabled through mutation. For each chromosome $k, k = 1, \dots, K$ in the population of generation g , a random probability measure $\alpha_k \in [0, 1]$ is generated. If α_k is greater than α , the probability of mutation, the chromosome k is subject to a random swap of two of its alleles as follows. Two random integers s_1 and s_1' (resp., s_2 and s_2') are randomly chosen from the discrete uniform $[1, r]$ (resp. $[1, c]$). The entries corresponding to $\hat{\pi}_{s_1, s_2, k}^g$ and $\hat{\pi}_{s_1', s_2', k}^g$ of k are then swapped. Mutation does not require the re-scaling of the alleles since the total relative frequency is fixed. The mutant replaces the least fit chromosome of the population if the former improves the latter. Once it completes the mutation step, GA ranks the population again.

To maintain the vitality of the population, GA culls the weakest chromosomes. Applying the survival of the fittest principle, GA selects the elite group consisting of the fittest K chromosomes of the mutated population. This group serves as the population of the next generation or iteration $g + 1$.

GA iterates through the above steps (i.e., crossover, mutation, and selection) until it satisfies a stopping criterion. Preliminary testing of the algorithm suggests that the stopping criterion should be a preset number of iterations $n_g = 1,000$. It ensures reasonably well-calibrated forecasts with a negligible fitness value of the best chromosome.

The above GA determines the joint probability distribution of two categorical variables X and Z based on an observed sample of paired observation. This distribution is used to determine the conditional probabilities of X given Z and of Z given X . However, the joint and conditional distributions are valid for a stationary point in time. In the following, the GA is extended to account for a time component (if applicable). Thus, GA will provide point probability forecasts for future or past points in time; allowing for the comparison of results of scientific studies undertaken at different points on the time horizon.

GA Across Time

For applications that involve time, GA is altered so that it evolves over time in a manner similar to a Markov chain. Let $t = t_1, t_2, t_3, \dots$, represent sequential points in time. At any arbitrary initial point in time t_i , the GA is executed as described above until a well-calibrated population, \mathbf{P}_{t_i} , comes to term. To move either forward or backward to instant t_i , the GA is executed once more using P_{t_i} as the initial population. The transition in time is made possible by altering the fitness function to

$$F_k^{t_i} = \frac{1}{N} \sum_{i=1}^r \sum_{j=1}^c (\hat{\pi}_{ij}^{t_i} - p_{ij}^{t_i})^2.$$

When applied forward (resp. backward) in time, this procedure sets t_i to t_{i+1} (resp. t_{i-1}). Time points do not need to be equally spaced on the time horizon. Explained in Figure 2 is the application of GA for transitions, where the present time is indicated via a dashed arrow and the future/past via a solid arrow. At the present time t_i , the initial population is generated randomly and GA is applied. The outcome of GA at the present time is then used as the initial population for the time t_i , regardless of whether $t_i = t_{i+1}$ or t_{i-1} .

GENETIC ALGORITHMS FOR CATEGORICAL DATA

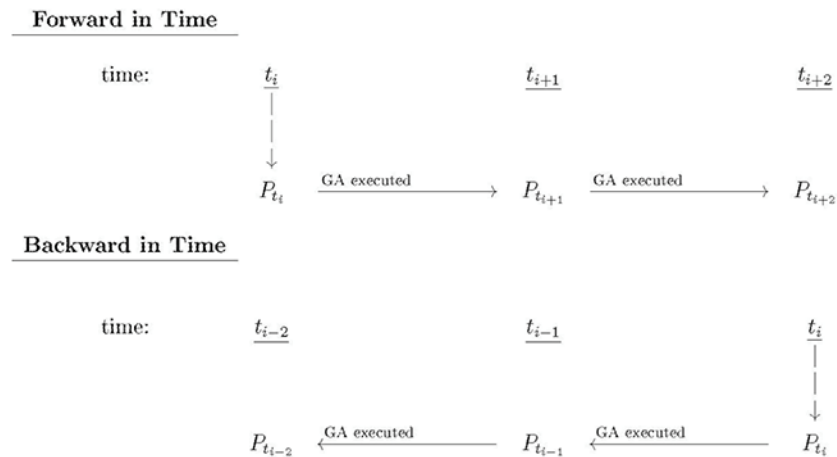


Figure 2. Forward and backward transition of GA in time.

A Cross-Calibration Application

In the assessment of stroke victims, standardized disability measures are commonly used. The scales are crucial in understanding the effectiveness of stroke treatments; yet, seldom is a patient assessed on more than one scale. A translation between two scales allows for the comparison among clinical trials and aids the development of alternative treatments.

Consider two commonly used standardized stroke disability measures, and apply GA cross calibration to form a feasible translation between them. The first is the Barthel Activity of Daily Living (ADL) Index (BI) attributed to Mahoney and Barthel (1965). It is a general measure of ADL, applied to a spectrum of medical conditions. The second is the Modified Rankin Outcome Scale (RS) (Rankin, 1957). It is a measure of the severity of disability in stroke victims. Currently, it is the most widely used measure of disability assessment for stroke victims (Saver et al., 2010). Much work has been done to compare the effectiveness of the measures and to determine whether the same clinical conclusion can be drawn from them (Sulter et al., 1999; Saver et al., 2010; Uyttenboogaart et al., 2007).

The BI defines 10 criteria of basic ADL and assesses the patients' capability to perform each of them. A minimum score of 0 is given if the patient is incapable of carrying out the task, and a maximum score is attributed if the patient can perform the ADL task independently. Partial scores, presented in increments of 5,

are allocated to patients who can perform the tasks, but with varying degrees of assistance. The scores of the 10 tasks are compiled to create an aggregate score with a maximum of 100. That is, a BI score of 100 indicates that the patient is physically independent.

The RS score assigns patients a discrete score from 0 to 5 depending on their degree of reliance on assistance and care. In contrast to the BI measure, a maximum RS score of 5 indicates the patient has severe disability and is highly dependent on nursing assistance. Whereas, a patient who exhibits no symptom of stroke debilitation and is independent is given a score of 0. Table 1 describes the 5 RS rankings and the 10 ADL criteria assessed by BI and their maximal achievable scores.

Table 1. The different measurement schemes: their measurement criteria and scores.

a. The BI criteria for ADL		b. The Modified Rankin Scale	
<i>Item</i>	<i>Maximum score</i>	<i>Item</i>	<i>Score</i>
Feeding	10	No symptoms	0
Transferring	15	No significant disability	1
Grooming	5	Slight disability	2
Toileting	10	Moderate disability	3
Bathing	5	Moderately severe disability	4
Walking	15	Severe disability	5
Stairs	10		
Dressing	10		
Bowel continence	10		
Bladder continence	10		

The data used in this example was taken from the Kansas City Stroke Study (KCSS), a prospective cohort study of 459 individuals designed to characterize the patterns of recovery of patients with mild, moderate, and severe stroke. As described by Duncan et al. (2000), the 459 individuals with stroke were assessed using both the BI and RS instrumentations 14 days after the incidence of stroke. A follow-up was performed at 1, 3, and 6 months after stroke. Table 2 summarizes the observed data.

All data was collected from hospitals in the Greater Kansas City area. The rating of the stroke patients in the study was performed on both the RS and BI scales by either a physical therapist or a study nurse. Despite the fact that the same enumerator rated each patient, the data is still subject to numerous sources of measurement error. One possible source is the two groups of raters: the study

GENETIC ALGORITHMS FOR CATEGORICAL DATA

nurses and the physical therapists. There can be differences both between and within these two groups on how they perceive and interpret the disability criteria measures. Likewise, a stroke patient's subjective interpretation of daily functions can vary widely from patient to patient depending on a wide spectrum of factors such as the patient's level of activity pre and post the advent of stroke. Another source of measurement error is how the enumerator perceives the patients' activity and the many interaction effects therein. All of these factors (among others) culminate adding noise to the observed sample distorting the true distribution of the data.

Table 2. Cross tabulation of the ADL scores of the KCSS at 1, 3, and 6 months after the onset of a stroke. The columns represent the RS score. The rows are the BI.

Month 1						Month 2						Month 3								
<i>B/RS</i>	0	1	2	3	4	5	<i>B/RS</i>	0	1	2	3	4	5	<i>B/RS</i>	0	1	2	3	4	5
0	0	0	0	0	1	10	0	0	0	0	0	1	7	0	0	0	0	0	0	5
5	0	0	0	0	0	9	5	0	0	0	0	0	2	5	0	0	0	0	1	2
10	0	0	0	0	3	1	10	0	0	0	0	0	2	10	0	0	0	0	1	1
15	0	0	0	0	2	1	15	0	0	0	0	5	0	15	0	0	0	0	3	0
20	0	0	0	0	5	3	20	0	0	0	0	3	0	20	0	0	0	0	3	2
25	0	0	0	0	7	1	25	0	0	0	0	7	0	25	0	0	0	0	4	1
30	0	0	0	0	7	0	30	0	0	0	0	6	0	30	0	0	0	0	4	0
35	0	0	0	0	8	0	35	0	0	0	0	7	0	35	0	0	0	1	3	0
40	0	0	0	2	14	0	40	0	0	0	0	3	0	40	0	0	0	0	4	0
45	0	0	0	0	4	0	45	0	0	0	0	3	0	45	0	0	0	0	2	0
50	0	0	0	1	8	0	50	0	0	0	1	6	0	50	0	0	0	2	3	0
55	0	0	0	1	9	0	55	0	0	0	0	5	0	55	0	0	0	3	5	0
60	0	0	1	5	9	0	60	0	0	0	3	6	1	60	0	0	0	3	4	0
65	0	0	1	5	3	0	65	0	0	0	4	3	0	65	0	0	1	0	5	0
70	0	0	1	12	3	0	70	0	0	1	11	8	0	70	0	0	0	6	1	0
75	0	0	1	19	6	0	75	0	0	0	5	0	0	75	0	0	0	12	2	0
80	0	0	1	18	3	0	80	0	0	6	12	0	0	80	0	0	0	9	0	0
85	0	0	4	26	0	0	85	0	2	4	21	0	0	85	0	0	3	18	0	0
90	1	0	7	24	1	0	90	1	2	9	23	0	0	90	0	3	11	13	0	0
95	1	4	31	13	0	0	95	1	4	24	20	0	0	95	2	6	35	16	0	0
100	2	17	62	11	0	0	100	7	44	72	9	0	0	100	11	57	62	11	0	0

Parmigiani et al. (2003) proposed a functional translation for the two measures using a statistical estimation approach. Although it produces adequate results, their approach requires that each characteristic of the relationship be modeled separately. GA avoids this. Its calibration accounts for all the relationship's characteristics intrinsically.

The objective is to determine the conditional probability distributions $P(BI|RS)$ and $P(RS|BI)$ at stationary time points and across time. Since both conditional distributions are functions of the joint distribution $P(BI,RS)$, GA determines only the latter. The GA is labeled vertical if applied at a stationary point in time and horizontal when applied either backward or forward across time.

Given in Table 3 are the joint distributions of RS and BI assessments at month 1. Table 3a is the result of a vertical GA at month 1 whereas Table 3b is the result of a backward GA starting at month 6 and moving in time to month 3 then to month 1. Both representations show good results; the negative correlation between the two scales is present, as expected, with higher probabilities attributed to the joint distribution of ratings along the counter diagonal in the lower triangle of Table 3.

Table 3. GA representations of the joint distributions after month 1 of the onset of a stroke. **a)** The joint distribution is independent of the information in months 3 and 6; **b)** The resulting joint distribution at month 1 when the GA is allowed to work backward in time from month 6 to month 3 to month 1.

a. Month 1: Random GA							b. Month 1: Time Reversal						
B/RS	0	1	2	3	4	5	B/RS	0	1	2	3	4	5
0	0.000	0.000	0.000	0.000	0.006	0.027	0	0.000	0.000	0.000	0.000	0.000	0.020
5	0.000	0.000	0.000	0.000	0.000	0.025	5	0.000	0.000	0.000	0.000	0.000	0.011
10	0.000	0.000	0.000	0.000	0.008	0.006	10	0.000	0.000	0.000	0.000	0.000	0.011
15	0.000	0.000	0.000	0.000	0.006	0.006	15	0.000	0.000	0.000	0.000	0.014	0.000
20	0.000	0.000	0.000	0.000	0.013	0.008	20	0.000	0.000	0.000	0.000	0.011	0.000
25	0.000	0.000	0.000	0.000	0.020	0.006	25	0.000	0.000	0.000	0.000	0.019	0.000
30	0.000	0.000	0.000	0.000	0.019	0.000	30	0.000	0.000	0.000	0.000	0.017	0.000
35	0.000	0.000	0.000	0.000	0.024	0.000	35	0.000	0.000	0.000	0.000	0.020	0.000
40	0.000	0.000	0.000	0.006	0.049	0.000	40	0.000	0.000	0.000	0.000	0.010	0.000
45	0.000	0.000	0.000	0.000	0.012	0.000	45	0.000	0.000	0.000	0.000	0.011	0.000
50	0.000	0.000	0.000	0.006	0.023	0.000	50	0.000	0.000	0.000	0.010	0.020	0.000
55	0.000	0.000	0.000	0.006	0.026	0.000	55	0.000	0.000	0.000	0.000	0.014	0.000
60	0.000	0.000	0.006	0.013	0.024	0.000	60	0.000	0.000	0.000	0.011	0.021	0.000
65	0.000	0.000	0.006	0.013	0.008	0.000	65	0.000	0.000	0.000	0.012	0.010	0.000
70	0.000	0.000	0.006	0.032	0.008	0.000	70	0.000	0.000	0.011	0.031	0.022	0.000
75	0.000	0.000	0.006	0.051	0.018	0.000	75	0.000	0.000	0.000	0.014	0.000	0.000
80	0.000	0.000	0.006	0.049	0.008	0.000	80	0.000	0.000	0.019	0.043	0.000	0.000
85	0.000	0.000	0.011	0.067	0.000	0.000	85	0.000	0.011	0.011	0.063	0.000	0.000
90	0.006	0.000	0.020	0.078	0.007	0.000	90	0.010	0.011	0.023	0.064	0.000	0.000
95	0.006	0.011	0.067	0.049	0.000	0.000	95	0.011	0.011	0.071	0.064	0.000	0.000
100	0.006	0.018	0.067	0.033	0.000	0.000	100	0.021	0.133	0.094	0.024	0.000	0.000

Similarly good results are reported for month 3, as depicted in Table 4, which gives its joint distribution. These results were achieved by applying the GA

GENETIC ALGORITHMS FOR CATEGORICAL DATA

forward (Table 4a), backward (Table 4b), and vertically (Table 4c); thus, allowing for the comparison of the three probability forecasts at month 3. All three GA approaches perform well, but the retrospective GA provides the best results. This conclusion is based on the smallest value of the fitness function and on how well the joint distribution exhibits the nature of the relationship between RS and BI.

Table 4. GA representations of the joint distributions at month 3 after stroke onset. **a)** The joint distribution resulting from the GA going back in time from month 6 to month 3; **b)** The GA results independent of the information in months 3 and 6; **c)** The results of the GA moving forward in time from month 1 to month 3.

a. Time Reversal							b. Random GA						
BIRS	0	1	2	3	4	5	BIRS	0	1	2	3	4	5
0	0.000	0.000	0.000	0.000	0.000	0.020	0	0.000	0.000	0.000	0.000	0.008	0.019
5	0.000	0.000	0.000	0.000	0.000	0.008	5	0.000	0.000	0.000	0.000	0.000	0.008
10	0.000	0.000	0.000	0.000	0.000	0.008	10	0.000	0.000	0.000	0.000	0.000	0.007
15	0.000	0.000	0.000	0.000	0.012	0.000	15	0.000	0.000	0.000	0.000	0.014	0.000
20	0.000	0.000	0.000	0.000	0.008	0.000	20	0.000	0.000	0.000	0.000	0.008	0.000
25	0.000	0.000	0.000	0.000	0.020	0.000	25	0.000	0.000	0.000	0.000	0.019	0.000
30	0.000	0.000	0.000	0.000	0.020	0.000	30	0.000	0.000	0.000	0.000	0.017	0.000
35	0.000	0.000	0.000	0.000	0.023	0.000	35	0.000	0.000	0.000	0.000	0.022	0.000
40	0.000	0.000	0.000	0.000	0.010	0.000	40	0.000	0.000	0.000	0.000	0.008	0.000
45	0.000	0.000	0.000	0.000	0.008	0.000	45	0.000	0.000	0.000	0.000	0.009	0.000
50	0.000	0.000	0.000	0.008	0.020	0.000	50	0.000	0.000	0.000	0.000	0.017	0.000
55	0.000	0.000	0.000	0.000	0.020	0.000	55	0.000	0.000	0.000	0.000	0.014	0.000
60	0.000	0.000	0.000	0.010	0.023	0.000	60	0.000	0.000	0.000	0.008	0.017	0.000
65	0.000	0.000	0.000	0.011	0.008	0.000	65	0.000	0.000	0.000	0.010	0.008	0.000
70	0.000	0.000	0.000	0.031	0.021	0.000	70	0.000	0.000	0.007	0.038	0.028	0.000
75	0.000	0.000	0.000	0.020	0.000	0.000	75	0.000	0.000	0.000	0.014	0.000	0.000
80	0.000	0.000	0.020	0.038	0.000	0.000	80	0.000	0.000	0.017	0.038	0.000	0.000
85	0.000	0.010	0.011	0.061	0.000	0.000	85	0.000	0.007	0.011	0.059	0.000	0.000
90	0.000	0.011	0.025	0.064	0.000	0.000	90	0.008	0.008	0.040	0.064	0.000	0.000
95	0.000	0.009	0.084	0.068	0.000	0.000	95	0.000	0.011	0.093	0.055	0.000	0.000
100	0.020	0.124	0.118	0.027	0.000	0.000	100	0.011	0.122	0.128	0.027	0.000	0.000

c. Forward in Time						
BIRS	0	1	2	3	4	5
0	0.000	0.000	0.000	0.000	0.006	0.022
5	0.000	0.000	0.000	0.000	0.000	0.007
10	0.000	0.000	0.000	0.000	0.000	0.007
15	0.000	0.000	0.000	0.000	0.025	0.000
20	0.000	0.000	0.000	0.000	0.008	0.000
25	0.000	0.000	0.000	0.000	0.020	0.000
30	0.000	0.000	0.000	0.000	0.019	0.000
35	0.000	0.000	0.000	0.000	0.020	0.000
40	0.000	0.000	0.000	0.000	0.007	0.000
45	0.000	0.000	0.000	0.000	0.009	0.000
50	0.000	0.000	0.000	0.006	0.021	0.000

Table 5 provides the joint distribution of RS and BI for month 6. Obtained in Table 5a is this joint distribution using the past information in month 1; a vertical GA is then applied for month 3 first then for month 6. Applied in Table 5b is a horizontal GA at month 6, using none of the data observed during months 1 and 6. Again, although both techniques show good results, the forward GA

produces slightly better results as it uses additional sample information for its forecast.

Table 5. GA representations of the joint distributions at month 6 after stroke onset. a. The resulting joint distribution at month 6 when the GA is allowed to move forward in time from month 1 to month 3 to month 6. b. The joint distribution independent of the information in months 1 and 3.

a. Month 6: Time Dependent							b. Month 6: Random GA						
B/RS	0	1	2	3	4	5	B/RS	0	1	2	3	4	5
0	0.000	0.000	0.000	0.000	0.006	0.027	0	0.000	0.000	0.000	0.000	0.000	0.020
5	0.000	0.000	0.000	0.000	0.000	0.025	5	0.000	0.000	0.000	0.000	0.000	0.011
10	0.000	0.000	0.000	0.000	0.008	0.006	10	0.000	0.000	0.000	0.000	0.000	0.011
15	0.000	0.000	0.000	0.000	0.006	0.006	15	0.000	0.000	0.000	0.000	0.014	0.000
20	0.000	0.000	0.000	0.000	0.013	0.008	20	0.000	0.000	0.000	0.000	0.011	0.000
25	0.000	0.000	0.000	0.000	0.020	0.006	25	0.000	0.000	0.000	0.000	0.019	0.000
30	0.000	0.000	0.000	0.000	0.019	0.000	30	0.000	0.000	0.000	0.000	0.017	0.000
35	0.000	0.000	0.000	0.000	0.024	0.000	35	0.000	0.000	0.000	0.000	0.020	0.000
40	0.000	0.000	0.000	0.006	0.049	0.000	40	0.000	0.000	0.000	0.000	0.010	0.000
45	0.000	0.000	0.000	0.000	0.012	0.000	45	0.000	0.000	0.000	0.000	0.011	0.000
50	0.000	0.000	0.000	0.006	0.023	0.000	50	0.000	0.000	0.000	0.010	0.020	0.000
55	0.000	0.000	0.000	0.006	0.026	0.000	55	0.000	0.000	0.000	0.000	0.014	0.000
60	0.000	0.000	0.006	0.013	0.024	0.000	60	0.000	0.000	0.000	0.011	0.021	0.000
65	0.000	0.000	0.006	0.013	0.008	0.000	65	0.000	0.000	0.000	0.012	0.010	0.000
70	0.000	0.000	0.006	0.032	0.008	0.000	70	0.000	0.000	0.011	0.031	0.022	0.000
75	0.000	0.000	0.006	0.051	0.018	0.000	75	0.000	0.000	0.000	0.014	0.000	0.000
80	0.000	0.000	0.006	0.049	0.008	0.000	80	0.000	0.000	0.019	0.043	0.000	0.000
85	0.000	0.000	0.011	0.067	0.000	0.000	85	0.000	0.011	0.011	0.063	0.000	0.000
90	0.006	0.000	0.020	0.078	0.007	0.000	90	0.010	0.011	0.023	0.064	0.000	0.000
95	0.006	0.011	0.067	0.049	0.000	0.000	95	0.011	0.011	0.071	0.064	0.000	0.000
100	0.006	0.018	0.067	0.033	0.000	0.000	100	0.021	0.133	0.094	0.024	0.000	0.000

In all executions of GA for this example, the fitness function converges quickly. Despite its small magnitude, the fitness function never converges to zero. This only reiterates the fact that the observed sample data used for the assessment of the chromosomes' fitness contains some noise, the sources of which were enumerated earlier. Figure 3 demonstrates how the fitness function decreases with the population evolution at a stationary point in time.

GENETIC ALGORITHMS FOR CATEGORICAL DATA

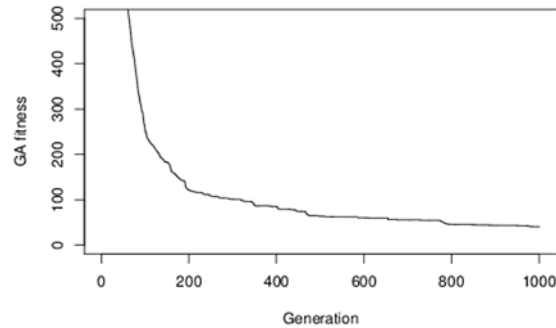


Figure 3. Convergence of the fitness function as the number of generations increases for a vertical GA applied at the stationary point 3 months after stroke onset.

In summary, GA performs well. The population converges quickly. In each generation, the chromosomes of the population display the negative correlations and properties that characterize the nature of the relationship between RS and BI. The time dependent GA performs better than the vertical GA because of the additional observed information being used. In particular, it is evident from the joint distribution of the first two time periods that a backward transition in time produces results that are more compliant with the expected nature of the relationship between the RS and BI measures.

Conclusion

Estimating the joint distribution of two categorical variables based on an observed sample data that contains some bias is an important topic and a cross-calibration problem. Because of its theoretical complexity and its widespread applications in several fields ranging from engineering to medicine to meteorology to population statistics. It is, herein, approximately solved using a non-traditional statistical method: genetic algorithm. Unlike other existing statistical methods, the adopted genetic algorithm does not make any assumption on the type or strength of the relationship between the categorical variables. It uses the observed sample to gauge the chromosomes of the successive populations. It converges rapidly to a good estimate of the true joint distribution. When applied over a time horizon, the genetic algorithm further enhances its estimates as it uses more observed data. When applied to the data collected for the Kansas City Stroke Study, it obtains logical point probability forecasts that concord with the true state of nature.

The proposed genetic algorithm based cross calibration approach can be tested with more sophisticated scoring rules or different fitness functions. Similarly, it can be applied to overcome missing data; in particular in clinical studies where subjects may move to different cities, die, or simply decide to stop participating in the study, and also in engineering set ups where the more reliable measurement methods are destructive or expensive.

References

- Ahn, J., Byun, H., Oh, K., and Kim, T. (2012). Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Systems with Applications*, 39(9), 8369–8379. doi: 10.1016/j.eswa.2012.01.183
- Aydilek, A. and Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, 233, 25–35. doi: 10.1016/j.ins.2013.01.021
- Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2
- Chen, R., Liang, C., Hong, W., and Gu, D. (2015). Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. *Applied Soft Computing*, 26, 435–443. doi: 10.1016/j.asoc.2014.10.022
- Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379), 605–610. doi: 10.1080/01621459.1982.10477856
- DeGroot, M. and Fienberg, S. (1983). The comparison and evaluation of forecasters. *The Statistician*, 32(1), 12–22. doi: 10.2307/2987588
- Duncan, P., Lai, S., and Keighley, J. (2000). Defining post-stroke recovery: implications for design and interpretation of drug trials. *Neuropharmacology*, 39(5), 835–841. doi: 10.1016/s0028-3908(00)00003-4
- Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1), 125–151. doi: 10.1146/annurev-statistics-062713-085831
- Huang, C. (2012). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, 12(2), 807–818. doi: 10.1016/j.asoc.2011.10.009

GENETIC ALGORITHMS FOR CATEGORICAL DATA

Huang, X., Zou, X., Zhao, J., Shi, J., Zhang, X., Li, Z., and Shen, L. (2014). Sensing the quality parameters of Chinese traditional Yao-meat by using a colorimetric sensor combined with genetic algorithm partial least squares regression. *Meat Science*, 98(2), 203–210. doi: 10.1016/j.meatsci.2014.05.033

Lichtenstein, S., Fischhoff, B., and Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky, Eds. *Judgment Under Uncertainty: Heuristics and Biases*, pp. 306–334. Cambridge, England: Cambridge University Press. doi: 10.1017/cbo9780511809477.023

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., and Wei, Y. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. *Mathematical and Computer Modelling*, 58(3–4), 458–465. doi: 10.1016/j.mcm.2011.11.021

Mahoney, F. and Barthel, D. (1965). Functional Evaluation: The Barthel Index. *Maryland Medical Journal*, 14, 61–65.

Murray, C. J. L., Tandon, A., Salomon, J., Mathers, C., and Sadana, R. (2002). Cross-population comparability of evidence for health policy. In C. J. L. Murray & D. B. Evans, Eds. *Health Systems Performance Assessment: Debates, Methods and Empiricism*, pp. 705–713. Geneva, Switzerland: World Health Organization.

Nieto, P., Fernandez, J., de Cos Juez, F., Lasheras, F., and Muniz, C. (2013). Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the Trasona reservoir (Northern Spain). *Environmental Research*, 122, 1–10. doi: 10.1016/j.envres.2013.01.001

Örkcü, H. H. (2013). Subset selection in multiple linear regression models: A hybrid of genetic and simulated annealing algorithms. *Applied Mathematics and Computation*, 219(23), 11018–11028. doi: 10.1016/j.amc.2013.05.016

Osborne, C. (1991). Statistical calibration: A review. *International Statistical Review*, 59(3), 309–336. doi: 10.2307/1403690

Parmigiani, G., Ashih, H., Samsa, G., Duncan, P. W., Lai, S., and Matchar, D. (2003). Cross-calibration of stroke disability measures. *Journal of the American Statistical Association*, 98(462), 273–281. doi: 10.1198/016214503000044

Rankin, J. (1957). Cerebral vascular accidents in patients over the age of 60: II. Prognosis. *Scottish Medical Journal*, 2(5), 200–215. doi: 10.1177/003693305700200504

Salomon, J., Tandon, A., and Murray, C. (2004). Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. *BMJ*, 328(7434), 258–263. doi: 10.1136/bmj.37963.691632.44

Saver, J., Filip, B., Hamilton, S., Yanes, A., Craig, S., Cho, M., Conwit, R., and Starkman, S. (2010). Improving the reliability of stroke disability grading in clinical trials and clinical practice: The Rankin Focused Assessment (RFA). *Stroke*, 41(5), 992–995. doi: 10.1161/strokeaha.109.571364

Sayed, H., Gabbar, H., and Miyazaki, S. (2009). A hybrid statistical genetic-based demand forecasting expert system. *Expert Systems with Applications*, 36(9), 11662–11670. doi: 10.1016/j.eswa.2009.03.014

Schervish, M., Seidenfeld, T., and Kadane, J. (2014). Dominating countably many forecasts. *The Annals of Statistics*, 42(2), 728–756. doi: 10.1214/14-aos1203

Stojanovic, B., Milivojevic, M., Ivanovic, M., Milivojevic, N., and Divac, D. (2013). Adaptive system for dam behavior modeling based on linear regression and genetic algorithms. *Advances in Engineering Software*, 65, 182–190. doi: 10.1016/j.advengsoft.2013.06.019

Sulter, G., Steen, C., and De Keyser, J. (1999). Use of the Barthel index and modified Rankin scale in acute stroke trials. *Stroke*, 30(8), 1538–1541. doi: 10.1161/01.str.30.8.1538

Uyttenboogaart, M., Luijckx, G., Vroomen, P., Stewart, R., and De Keyser, J. (2007). Measuring disability in stroke: relationship between the modified Rankin scale and the Barthel index. *Journal of Neurology*, 254(8), 1113–1117. doi: 10.1007/s00415-007-0646-0

van Buuren, S. and Hopman-Rock, M. (2001). Revision of the ICIDH severity of disabilities scale by data linking and item response theory. *Statistics in Medicine*, 20(7), 1061–1076. doi: 10.1002/sim.723

van Buuren, S., Eyres, S., Tennant, A., and Hopman-Rock, M. (2001). *Response conversion: A New Technology for Comparing Existing Health Information*. TNO Prevention and Health, Division Public Health.

Vitkovský, J., Simpson, A., and Lambert, M. (2000). Leak detection and calibration using transients and genetic algorithms. *Journal of Water Resources Planning and Management*, 126(4), 262–265. doi: 10.1061/(asce)0733-9496(2000)126:4(262)

Wibowo, A. and Desa, M. (2012). Kernel based regression and genetic algorithms for estimating cutting conditions of surface roughness in end milling

GENETIC ALGORITHMS FOR CATEGORICAL DATA

machining process. *Expert Systems with Applications*, 39(14), 11634–11641. doi:
10.1016/j.eswa.2012.04.004