33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

A Data-Driven Process for the Development of an Eyes-closed EEG Normative Database

Monica Aguilar, Marco Congedo and Javier Minguez

Abstract— **A normative database constitutes a representative sample of a neurologically and clinically healthy population. The practical utility of a normative EEG database is to evaluate the clinical status of a subject whose EEG patterns statistically diverge from average population patterns. These normative data are daily used in clinical practice and in the evaluation of therapeutical interventions. The main obstacle of all normative databases developed to date is inter-individual variability. Such difficulty has been addressed by stratifying the population by age and then using regression in the EEG groups to bound variability, which is always an approximation. This paper describes the first data-driven EEG normative database that explicitly deals with EEG variability by stratifying the population based on their EEG patterns. The database has been constructed for 84 subjects in eyes-closed condition and has been validated by cross validation, leading to a global specificity of 100%.**

I. INTRODUCTION

Diagnosis and treatment of neural pathologies have been influenced by modern clinical science developments. In recent decades, EEG has gained popularity due to its capacity of improving clinical diagnostic through the use of normative databases. A normative database is established through the collection of EEGs from a healthy population. The database is obtained under the same recording conditions, which constitutes a representative sample of a neurologically and clinically healthy population. The practical utility of a normative EEG database is the evaluation of the neurological status of a patient, with the purpose of establishing a clinical diagnosis. Such diagnosis is obtained through the comparison of the subject with a population of healthy individuals, in order to identify atypical features and the magnitudes of deviation. In practice, EEG is used for the diagnosis of pathologies such as epilepsy, ADHD, stroke, etc; as well as for the evaluation of the course and outcome of a therapy (by measuring whether EEG evolution occurs towards the normative EEG). Normative databases are becoming so widely accepted that quality standards have been established by the American EEG Association [1].

The first normative database was developed in the 1950s at UCLA [2], followed by the development of other bases [3], [4], [5]. Currently, there are several available normative EEG databases that may have relevance for clinical diagnosis and therapy evaluation. However, the usefulness extent of these databases is greatly determined by the degree of open disclosure of its contents [6], [7], [8] (open disclosure of the number of subjects per age group, gender, sample demographics, geographic location of the samples, quality control measures, acquisition and technical procedures, etc.).

The main obstacle in the development of normative databases is the inter-individual variability of the EEG, as in principle the database represents invariant features of healthy population. For this reason a population is usually stratified by age, in order to bound the effect of such variability (inter-subject) [3], [10], [11]. Once the base is stratified, either there is no special processing for the group [4], [6] or a least-squares regression is used to fit a function of the EEG data samples over the entire age range of the subjects [12]. In both cases, representative EEG patterns are obtained for each group age. The general drawback of these designs is that although EEG has usually low variability in adults, it is still influenced by factors such as rate of maturation or scalp-to-skull conductivity. These factors have great impact specially in young populations due to EEG variability, and thus the maturational lag in cortical development hinders the identification of a representative group age.

This paper describes the first data-driven normative EEG database that explicitly addresses EEG variability. To this end, the population is stratified based on their EEG patterns, which represents an alternative to existing normative databases. The data-driven design consists firstly of a consistent and stochastic optimization process, achieving gaussianity in the EEG power spectrum of each subject, and secondly of an unsupervised clustering method that stratifies the EEG based on the similarity of the EEG power distributions of the subjects. The base has been constructed for 84 subjects in eyes-closed condition and has been validated by cross validation, leading to a global specificity of 100%.

II. METHODS

A. Data Recording

The EEG data was provided by Nova Tech EEG Inc., Mesa, AZ. The population sample was composed of 84 healthy adults, aged between 18 and 30 years. Exclusion criteria included psychiatric history of drug/alcohol abuse in any relative and participant, head injury (at any age, even very young), headache episodes, physical disability, and epilepsy. EEG were recorded during 3-5 minutes while the subjects sat with eyes closed on a comfortable chair in a quiet and dimly-lit room. EEG data were acquired at the 19 standard leads prescribed by the 10-20 international system

M. Aguilar and J. Minguez are with the Aragon Institute of Engineering Research (I3A) of the University of Zaragoza. M. Aguilar is also with Bit&Brain Technologies S.L. E-mail: monicaaguilar@bitbrain.es, jminguez@unizar.es. M. Congedo is with the French National Center for Scientific Research (CNRS), France. E-mail: marco.congedo@gmail.com. This work has been partially supported by Spanish projects HYPER-CSD2009-00067 and DPI2009-14732-C02-01.

(FP1, FP2, F7, F3, FZ, F4, F8, T3, C3, CZ, C4, T4, T5, P3, PZ, P4, T6, O1, O2) using both earlobes as reference, and enabling a 60 Hz notch filter to suppress power line contamination. The impedance of all electrodes was kept below 5K Ohms. Data were acquired using a 12-bit A/D NeuroSearch-24 acquisition system (Lexicor Medical technology, Inc., Boulder, CO) and sampled at 128 Hz. In order to minimize inter-subject variability, all biological, instrumental and environmental artifacts were removed from data, paying particular attention to biological artifacts generated by eyes, hearth, and muscles of the neck, face and jaw.

Power spectral density was estimated through Fast Fourier Transform (FFT) for each 1 second epoch, in each recording channel, using a Tukey window, yielding a 0.06 Hz frequency resolution in the 2-30 Hz using zero-padding. Band powers were defined as delta (2-3.5 Hz), theta (3.5-7.0 Hz), alpha (7.0-13.0 Hz) and beta (13.0-30.0 Hz). Absolute band power was calculated from the area under the curve of the power spectrum, using a trapezoidal interpolation between the two limiting frequencies for the four bands.

B. Normative Database Development

The design of the EEG normative database consists of three steps: (a) obtainment of reliable data to provide consistency in the EEG acquisition process; (b) achievement of gaussianity in the EEG power spectrum of the subjects; and (c) stratification of the subjects based on power spectrum distributions. In addition, the development of an EEG normative database includes the necessary mathematical tools to verify whether the EEG of new subjects is within normality and to measure the degree of deviation.

1) Step 1: Reliability: The intra-individual variability of the EEG was verified through Test-Retest and Split & Half, using in both cases the Spearman and Pearson correlation. The results obtained did not achieve a minimum reliability of 0.8 (minimum in accordance with normative databases [6]), possibly due to a reduced length of data imposed by a restriction in the acquisition protocol. For this reason, a new technique to improve EEG reliability was developed, where new epochs were created by mergence. This new method divides the original EEG in half epochs and then joins the half-epochs randomly, obtaining *n-1* new epochs (where *n* is the number of epochs in the original EEG), leading to a new set of *n+(n-1)*. The process is repeated until a minimum reliability of 0.8 is obtained through Split & Half in all power bands. Note that this method iteratively improves EEG reliability, while the EEG spectrum is only modified by a high frequency artifact artificially introduced due to the discontinuity in the merging point of the epochs. There is no influence on the frequency range used in the EEG analysis (2-30Hz). The mean value of reliability obtained across all subjects was 0.894. Figure 1 displays the reliability per subject as well as the original and new number of epochs.

2) Step 2. Gaussian Filter: The second step uses filters to convert marginal power distributions in all bands and channels into Gaussian distributions [19]. This occurs because absolute band powers are non-Gaussian and their probability density function is characterized by a positive skew. The filtering process is constituted by two stages: (a) a non-linear parametric transformation of all power distributions in all channels to obtain distributions closer to the Gaussian distribution; and (b) application of a consistent and stochastic optimization technique to simultaneously filter marginal distributions of all the power bands and channels until Gaussian distributions are followed.

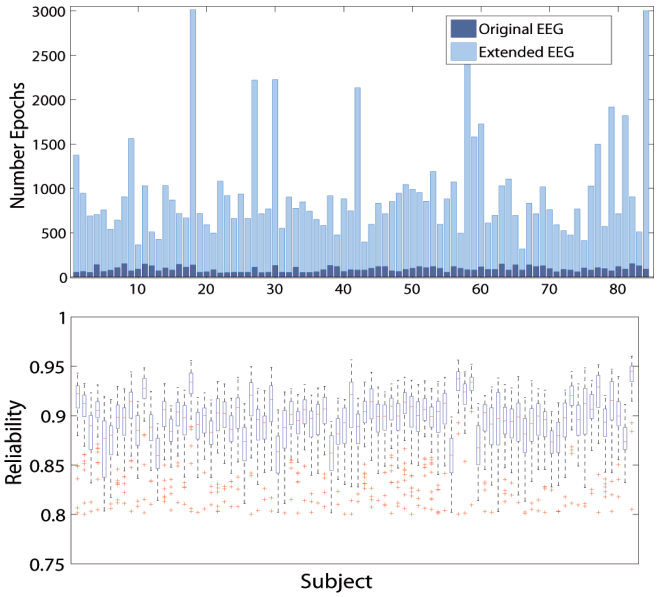

Fig. 1. Upper figure: Number of epochs before and after extension of process for every subject. Lower figure: Mean reliability for every subject across the marginal absolute power distributions.

The first stage of filtering is transformation. In EEG-normative literature, it is usual to apply transformations such as log(x), log(x + 1), $\sqrt{(x)}$, $\sqrt[3]{(x)}$, or $1/\sqrt{(x)}$ [13], [14]. However, statistical literature describes a range of more systematic methods for transforming distributions towards Gaussian distributions, such as the Box-Cox transformation [15]. Box-Cox is a parametric transformation, based on a wider family of transformations, with a simple procedure to build the best transformation based on data (it is a data-driven transformation). As the Box-Cox is an uni-dimensional transformation, there is a necessity to find the best transformation to jointly transform all marginal distributions of power bands in all channels and subjects. A mean parameter of each marginal transformation is obtained across the subjects of the database. Note that this transformation is fully dependent on the set of subjects of the base. The following table shows the number of marginal distributions achieving gaussianity, averaged across the subjects from Channels$\times$Bands$= 19 \times 4 = 76$ marginal distributions (the Anderson Darling test was used to verify normality [16]).

	No Transf	$1/\sqrt{x}$	$\sqrt{x}$	log(x)	Box-Cox
% Gaussian distrib.	0	1.02	6.92	28.10	48.50

Table I: Comparison of transformations

The second stage is to use a stochastic optimization

technique to simultaneously filter the marginal distributions per subject until all follow Gaussian distributions. A genetic algorithm was designed to find the subset of epochs (from the original set of EEG epochs per subject) that optimize an objective function, which is proportional to the significance of a gaussianity test applied to all marginal distributions of power bands and channels (after the Box-Cox transformation). The value of this objective function is maximal when all distributions in band powers and channels are Gaussian for a given interval of confidence (the Anderson Darling test was used to verify normality [16] with p-value=0.05). The genetic algorithm was implemented in such a way that the solution space could be efficiently explored. The search space has Channel×Band=$19 \times 4 = 76$ dimensions and the optimization landscape is highly non-linear (leading to a difficult computational problem).

Consistency of the stochastic optimization technique was tested though repetitive filterings. The resulting distributions from different repetitions were compared through the Kolmogorov-Smirnov (K-S) test, to check whether the distributions were the drawn from the same distribution. Additionally, these repetitive computations were used to check the effect of the filter size (% of epochs filtered from the original data set). After an intense computational exercise, 63 subjects of the 84 adults achieved optimization convergence. Consistency of the solution was 100% as for all subjects the K-S tests always accepted the null hypothesis (distributions resulting from repetitive optimizations always originate from the same continuous distribution).

3) Step 3. Unsupervised Clusterization: The third step is to stratify the subjects based on their power spectrum distributions. This problem can be stated as an unsupervised clusterization problem driven by similarities in the power distribution of the subjects. There is no explicit knowledge of the number of clusters, of which information will be used to build the clusters, or to which cluster a subject will belong. Note that this is a data-driven strategy that will find the stratification of the database based on data itself, not on subjective knowledge such as the age of the subjects.

A hierarchical agglomerative clustering was chosen from among other techniques such as K-means or Gaussian Mixture Models [17]. Hierarchical agglomerative clustering was chosen because the feature space presents large dimensionality (76 dimensions) and the application of other techniques resulted in non reliable solutions. Note that a dimension reduction with standard techniques such as PCA was avoided due to the impossibility of finding a criterion to valid information loss when other subjects were tested against the database (following the same data process). Given the power distributions of the subjects, a similarity measurement had to be provided in order to apply the unsupervised technique. The Bhattacharyya distance [18] was used because it measures the distance between distributions and has a closed-form solution for Gaussian distributions (Bhattacharyya measures the overlapping proportion in the histograms of two distributions). The agglomerative clustering can be constructed by using any similarity measurement based on the power distributions. Two measurements were proposed accounting for global and local aspects of distributions. The first measurement $\overline{DB}_T(s_p, s_t)$ is a weighted similarity for all distributions, and the second measurement $\overline{DB}_{\theta/\beta}(s_p, s_t)$ weights only the theta-beta ratio distribution in central electrodes (typical phenotype used in the diagnosis of attention deficit disorder [20]):

$$\overline{DB}_T(s_p, s_t) = \sqrt{\sum_{i=1}^{nChan} \sum_{j \in \{\delta,\theta,\alpha,\beta\}} DB_{i,j}(s_p, s_t)^2} \quad (1)$$

$$\overline{DB}_{\theta/\beta}(s_p, s_t) = \sqrt{\sum_{i=1}^{nChanCentral} DB_{i,\theta/\beta}(s_p, s_t)^2} \quad (2)$$

where DB_{ij} is the Bhattacharrya distance between distributions for the absolute band power in channel i and frequency band j of subjects s_p and s_t.

The agglomerative clustering with these similarity measurements works iteratively, where the two most similar clusters of subjects in terms of the previous distance are merged in every iteration. In this process, the distance between clusters is calculated from the similarity of their farthest members. The result of the clustering process is shown in a dendrogram (the x-axis is the set of subjects, the y-axis is the similarity measurement of clusters, and the tree structure shows the clustering). The final step is to establish a similarity threshold in order to set the result of the cluster in the hierarchy. To this end, the knee-rule was used over the objective function $DBC(k)$, where $DBC(k)$ is the maximum distance between clusters with the partition of *k* clusters (adapted to the similarity measurement used, see above Equations). This function is a monotonically decreasing function of *k* and the knee is the point with a significant change of slope DBC_{thr}. Figure 2 shows the result of the agglomerative clustering (stratification of database) for the two proposed measurements.

4) Step 4. Associated tools: The final objective of the normative database is to compare a new individual against normative data to find dis/similarities in global or specific features of the EEG. Let be $S^i = \{s_1^i, \ldots, s_p^i\}$ one of the m clusters computed with p subjects (with $i = 1 \ldots m$). Let be DBC_{thr} the threshold computed. A new subject s_{new} is in of the normality if:

$$\forall_i \overline{DB}_T(s_{new}, S_i) < DBC_{thr} \quad (3)$$

If subject s_{new} is within the normality, the most similar set of subjects naturally is given by the distance to the closer cluster. However, if the subject is out of the normality, this distance presents the minimum dissimilarity that could also be used to measure progress within a therapeutic process. Note that a similar expression is given for the other metric $\overline{DB}_{\theta/\beta}$.

C. Validation

Cross-validation is used to assess how the results of the clustering process generalize to an independent data set. Two

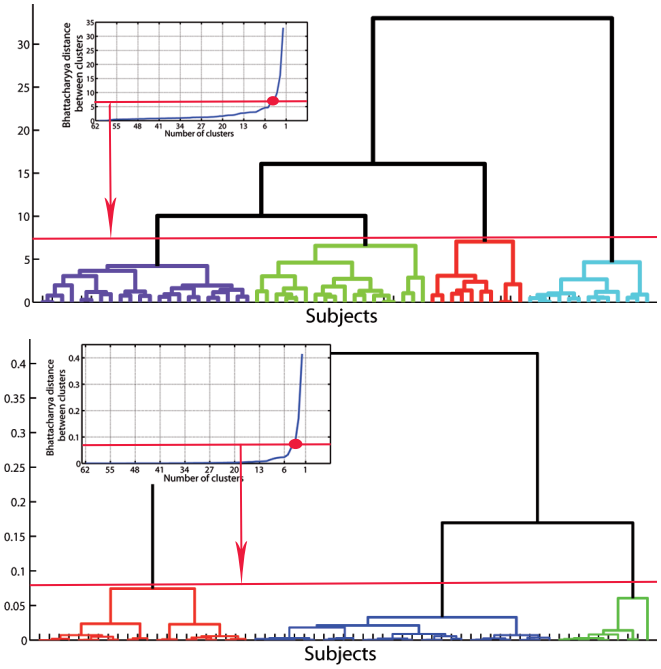

Fig. 2. Stratification of the EEG normative database with agglomerative clustering for (Upper part) the global similarity measurement and (Lower part) the similarity measurement of θ/β distribution ratio

different measurements are calculated: global specificity and internal specificity, both with a leave-one-out cross validation strategy using the EEG database subjects.

1) Global specificity: Global specificity is determined through the comparison of individual subjects against the normative database built without the specific subject, in order to verify whether the EEG is within the normality defined by the database. Global specificity is given by $Specificity = \frac{TN}{TN+FP}$, which determines the proportion of healthy subjects who are correctly labeled as normal by the test. Global specificity is computed through cross validation with leave-one-out (i.e., testing each subject against the normative database built with the rest of subjects). The TN obtained was 63 and $TN + FP$ are the 63 subjects that constitute the normative database. Therefore, a 100% global specificity was obtained for the two defined metrics.

2) Internal specificity: Internal specificity is computed through cross-validation, to check whether the TN obtained are correctly classified in the original cluster. Table 2 and Table 3 show the confusion matrix where the local specificity is given for the four clusters in both cases (global similarity metrics and θ/β distributions).

	Original Test			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	86.36	27.78	10.00	0
Cluster 2	4.55	61.11	0	0
Cluster 3	9.09	0	90.00	0
Cluster 4	0	11.11	0	100

Table 2: Local specificity for every original cluster for all 76 EEG features.

	Original Test		
	Cluster 1	Cluster 2	Cluster 3
Cluster 1	68.19	0	0
Cluster 2	31.81	93.55	10.00
Cluster 3	0	6.45	90.00

Table 3: Local specificity for every original cluster in ratio Theta/Beta.

III. CONCLUSIONS

This work designed and developed the first data-driven normative database, which provided a stratification of subjects based explicitly on EEG patterns. These EEG patterns can be in the form of global measurements or specific phenotypes with clinical relevance. The process was validated by cross validation with a 100% in specificity.

REFERENCES

[1] F. Duffy, J.R. Hughes, F. Miranda, P. Bernad, P. Cook. "Status of quantitative EEG (QEEG) in clinical practice", *Clinical Electroencephalography*, 25(4) 1994 ,vi-vixxii.

[2] W. R. Adey, "Data acquisition and analysis techniques in a Brain Research Insitute" *Ann NY Acad Sci* 31(115), 844-866

[3] M. Matousek, I. Petersen "Automatic evaluation of background activity by means of age-dependent EEG quotients" *EEG & Clin. Neurophysiol.*, 35, 603-612, 1973.

[4] E.R. John, L.S. Prichep, P. Easton "Normative data banks and neurometrics: Basic concepts, methods and results of norm construction." *Handbook of electroencephalography and clinical neurophysiology: III. Computer analysis of the EEG and other neurophysiological signals* pp.449-495, 1987. Amsterdam: Elsevier.

[5] M.A.T. Figueiredo, A. K. Jain (2002), "Unsupervised Learning of Finite Mixture Models",*IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 381-396.

[6] M .Matousek, I. Petersen "Frequency analysis of the EEG background activity by means of age-dependent EEG quotientes" *Automation of clinical electroencephalography(P.Kellaway and I. Petersen, eds)*,1973

[7] R.W. Thatcher "EEG normative databases and EEG biofeedback" *Journal of Neurotherapy*,1998, 2(4), 8-39.

[8] R.W. Thatcher, R.A. Walker, C. Biver, D. North, R. Curtin "Quantitative EEG Normative databases: Validation and Clinical Correlation" *Journal Neurophysiology*, 112, 1729-1745.

[9] E.R. John "Neurometrics: Quantitative Electrophysiological Analyses" *Functional Neuroscience*, 1977, New Jersey: L. Erlbaum Assoc.

[10] R.W. Thatcher, R.A. Walker, C. Biver, D. North, R. Curtin "Quantitative EEG Normative databases: Validation and Clinical Correlation" *J. Neurotherapy*, 2003, 7, 87122.

[11] R.W. Thatcher, D. North, C. Biver "EEG inverse solutions and parametric vs. non-parametric statistics of Low resolution Electromagnetic tomography (Loreta)" *Clin. EEG and Neuroscience*, 2005, 36(1), 19.

[12] E. R. John, H. Ahn, L.S. Prichep, M. Trepetin, D. Brown, H. Kaye "Developmental equations for the electroencephalogram" *Science*, 1980, 210, 12551258.

[13] T. Gasser, P. Bacher, J. Macks "Transformations towards the normal distribution of broad band spectral parameters of the EEG" *Electroencephalogr Clin Neurophysiol* 1982, 53, 119-24.

[14] WP. Dunlap, RS. Chen, T. Greer "Skew reduces test-restest reliability" *Journal Appl Physchol* 1994, 79, 310-13.

[15] GEP. Box, DR. Cox "An analysis of transformations" *J R Stat Soc B* 1964, 26, 211-52.

[16] Jr. Thode "Testing for Normality" 2002, *Marcel Dekker, New York.*

[17] J.A. Hartigan "Clustering Algorithms", (1975), Wiley.

[18] A. Bhattacharyya "On a measure of divergence between two statistical populations defined by probability distributions" *Bull. Calcutta Math. Soc.*, 35 (1943) pp. 99109.

[19] M. Congedo, J.F. Lubar "Parametric and Non-Parametric Normative Database Comparisons", *Electroencephalography: A Simulation Study on Accuracy, Journal of Neurotherapy*,(2003) 7(3/4), 1-29.

[20] Lubar, J. F. "Discourse on the development of EEG diagnostics and biofeedback treatment for attention deficit/hyperactivity disorders", *Biofeedback and Self-Regulation*,1991, 16, 201-225.