World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:16, No:10, 2022

Designing Social Care Policies in the Long Term: A Study Using Regression, Clustering and Backpropagation Neural Nets

Sotirios Raptis

Abstract—Linking social needs to social classes using different criteria may lead to social services misuse. The paper discusses using ML and Neural Networks (NNs) in linking public services in Scotland in the long term and advocates, this can result in a reduction of the services cost connecting resources needed in groups for similar services. The paper combines typical regression models with clustering and cross-correlation as complementary constituents to predict the demand. Insurance companies and public policymakers can pack linked services such as those offered to the elderly or to low-income people in the longer term. The work is based on public data from 22 services offered by Public Health Services (PHS) Scotland and from the Scottish Government (SG) from 1981 to 2019 that are broken into 110 years series called factors and uses Linear Regression (LR), Autoregression (ARMA) and 3 types of back-propagation (BP) Neural Networks (BPNN) to link them under specific conditions. Relationships found were between smoking-related healthcare provision, mental health-related health services, and epidemiological weight in Primary 1(Education) Body Mass Index (BMI) in children. Primary component analysis (PCA) found 11 significant factors while C-Means (CM) clustering gave 5 major factors clusters.

Keywords—Probability, cohorts, data frames, services, prediction.

I. Introduction

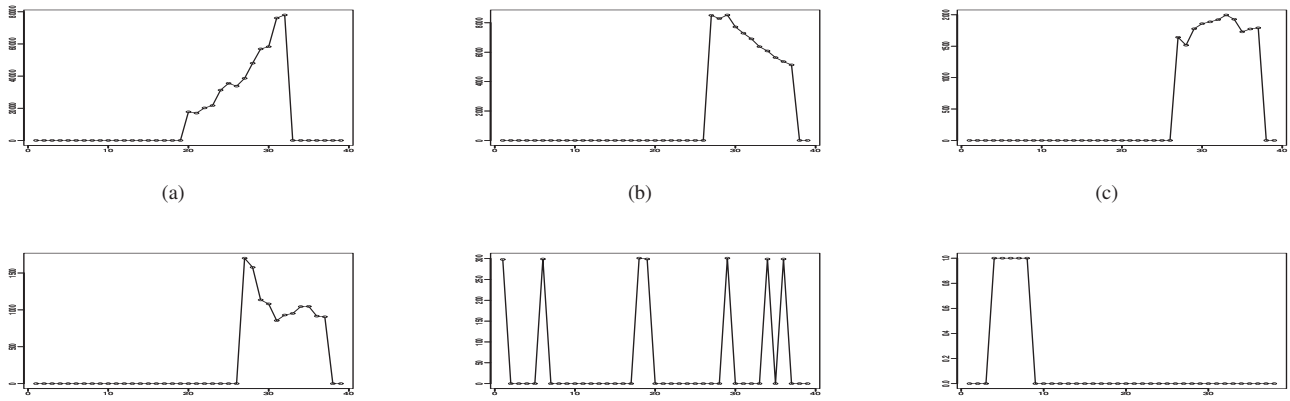
THERE have been concerns about the system of publicly funded social care in England and in Scotland for more than 20 years. The generic belief is that a good prediction of the demand helps with better managing servicesas resources. The paper here advocates this can be achieved by classifying servicesinto cohorts and can create additional revenue that in turn can be used to face the growing demand. Saving on resources can be based on connecting services using classification and prediction. Healthcare cost is forecasted, in the UK, and by 2031 to reach the range of several Bn pounds a year, if the cost is not better managed. As an indication of that [1] discusses that the cost can become as high as 12 billion by 2030/31 at an average rate of 3.7 percent a year. The present paper attempts to address this problem using public H&Sc data available on PHS' website [2] and from the SG [3] posted by June, 2019. The data used here were counts of patients (called 'Value' attribute in the data) and containing the parameters for each service. PCA was applied to see the most important ones after normalization was applied as discussed in [4]. Works that mine services sequences (patterns) belong to the same category as they use similarity metrics to patterns that are stored in a database [5] that is based on prediction or on being in the same cohort. Zero padding was used for data imputation in case of missing data and relevant methods can be found in [6] and for imputation using Markov models in [7], [8] that use statistical models to approach the missing data. Linear prediction for example, Auto-regressive Moving Average (ARMA) is discussed in [9] while the linear association of different service parameters is also question in [10] and a review of linear methods in healthcare (HC) is presented in [11]. The paper is organized as follows. In the 1^{st} Section, the nature of the data processed is better explained and the main analysis is given by introducing the LR and its application to PHS data. The ARMA/AR prediction is presented along with cross-correlation (CC), C-Means (CM), and PCA. Indicative comparisons and results are presented in the 2^{nd} Section using both the classification and the regression methods and are accompanied by comparative co-plots or tabular forms when numerical comparisons. Emphasis is given to the probabilistic association of H&Sc factors and to the notion of error measures such as (coefficient of determination), Median Relative Error (MRE), and Median Absolute Error (MAE).

Mr. Raptis is with the School of Design and Informatics, Abertay University, Dundee, Scotland, Bell Street, Dundee, DD1 1HG, with MacMillan Cancer Support UK, Caledonian Exchange, 19A Canning St, Edinburgh EH3 8EG (current collaborator), and with National Health Services (NHS) Scotland, Gyle Square, 1 South Gyle Crescent, Edinburgh, EH12 9EB, Scotland, UK (during the works of the paper) (e-mail: sotnraptis@yahoo.com).

II. Materials and Methods

The raw data processed can be visualised in Fig. 1 (a) where we see the plot of the service (S1.A2.L2) (*'Alcohol use among young people.Age.15'*), in Fig. 1 (b) the plot of (S21) (*'Drug use among 13 and 15 year olds in Scotland) . main client group in care home . other groups (no acronym shown)'*), in Fig. 1 (c) (S19. A2. L1) (*'% of children classed healthy weight, overweight, obese, severely obese at Primary 1 review . Age . 13'*), in Fig. 1 (d) (S7 . A6. L4) (*'Number of single rooms in care homes . type of tenure . owned outright'*), in Fig. 1 (c) (S8) (*'Drug related hospital discharges). value'*) , in Fig. 1 (f) the plot of (S16 . A2. L2) (*'Occupancy rate in care homes by type of provision . Age . 15'*).

Most representative and varying H&Sc factors across the H&Scs groups are shown from Figs. 1 (a)-(f). In those figures the Y-axis shows the attendances, L's are the levels, some are shown in Table. I. The boldfaced items represent attributesnames, and the rows(items) below represent the levels



(a) (b) (c)

(d) (e) (f)

Fig. 1 Representative services' plots

TABLE I
SERVICES BREAKDOWN INTO ATTRIBUTES AND LEVELS

Number of home care clients by care type or disability (S4)	**Headcount of general practice workforce**(S3)	**Living arrangements for home care clients**(A4)	**Home care clients**(S22)	**Repeated emergency admissions**(S2)	**Number, percent, for low birth-weight (¡=2.5Kg)**(S10)	**Smoking behaviour**(A5)	**Type of tenure**(A6)
Age	Value(no attributes)	2 or more clients in home	Age	Value(no attributes)	Single births	Non-smoker	All
Gender		alone	Gender		Birth weight	Occasional smoker	Owned mortgage loan
Home care client group		other living conditions			Age bands	Regular smoker	Owned outright
							Rented
Self assessed general health(A9)	**Weight category**(A7)	**Household type**(A8)	**Birth weight**(A10)	**Gender**(A3)	**Age**(S2)	**Age bands**(A1)	
Bad	Epidemiological	Adults	singleton Births	Male	13	18 years and over	
Fair	Healthy weight	All	Low weight births	Female	15	18-74 years	
Good	Epidemiological $*1$	Pensioners			All	75 years and over	
Very bad	Epidemiological $*2$						
Very good	Epidemiological $*3$						
	Very good						

$*1$ Obese, $*2$ Overweight, $*3$ Underweight

of the attributes. Where there are mo specific attributes, as in services(S3), the *'value'* attributeis used and it means all counts are considered. The X-axis shows the span of 39 years. The breakdown of the service packs into their attributes and levels is shown in Table I. The services names and acronyms are listed in Table II. The groups are represented as parts of linear prediction equations. Their full names are shown, their acronyms(inside parentheses as (*'S.A.Z'*) triplets, their dates of recording and the numbers of characteristics(attributes) that accompany them.

The attributes and their levels are part of linear relationships. Some are presented in Table III. Some attributes like ages, gender are shared among H&Sc groups. The table shows the LR probabilities $P1$,$P2$ and $P3$, the error metrics $RMSE \rightarrow Er1$, $MAE \rightarrow Er2$, $MRE \rightarrow Er3$ across the 39 years. The naming of the factors follows the attributeslevels and H&Sc names convention defined in the Tables 1 and 2. The NNs are also compared for the 3 NN structures with (3,10,15) nodes in the (H) hidden layer.

Fig. 1 shows indicative breakdowns of H&Sc data frames (the (S)s) into their attributes (the A's) and of their attributesinto their levels (items in "A"s lists).

In Fig. 2 (g) the plots show the error falls as we add more cross-correlated H&Sc factor. Same results were obtained by computing the clusters before with K-Means that are not shown. In Fig. 2 (g) LR, ARMA errors compared for H&Sc factor S2.Value using few AR lags (1 to 6) and LRorders and RMSE, Median Absolute Error and Median Relative Error errors. In Fig. 2 (c) we see the LR-CM prediction with variable orders original, 1^{st} and 7^{th} normalized shown, and selected years (1981:1991) for Hospital Admissionof all ages and genders and admission reason *'Elective Planned'* for *le.r*= 0.3.

In Fig. 3 the pattern for the X-axes {*'N'*,'/ others' } means that the 1^{st} independent factor is drawn sequentially from the set {'Delayed discharges (DD's) . monthly census . ratio', 'Care home clients . gender(Male)', 'Alcohol use among young people . age(13-15)', 'Alcohol use among young people . age (15-18)'} while the rest of the independent's are the left *'N'*'s.

Plots for indicative linear connections found ($HScfactor_i = LR(HScfactor_j), i <> j$) are shown

TABLE II
H&Sc SERVICES GROUPS INCLUDING FULL NAMES, DATES, ACRONYMS

H&Scs full names	A^a	B^b	C^c	**H&Scs full names**	A^a	B^b	C^c
1. Alcohol use among young people	S1	1998-2010	3	2. Repeated emergency admissions	S2	1998-2010	1
3. Headcount of General Practice Workforce	S3	2008-2019	1	4. Number of home care clients by care type or disability	S4	2005-2009	2
5. Living Arrangements for Home Care Clients	S5	2005-2009	3	6. Intensive Home Care	S6	2002-2011	1
7. Number of single Rooms in care homes	S7	2007-2017	2	8. Drug related hospital discharges	S8	1996-2018	1
9. Home care services	S9	2005-2009	2	10. Number, percent, for low birth weight (<=2.5Kg) for single births	S10	2000-2019	2
11. Mental wellbeing by tenure, household type, age, sex, disability	S11	2014-2017	4	12. Number of general practices (GPs) with registered patients	S12	2007-2019	1
13. Number of care homes by type of provision	S13	2007-2017	6	14. Occupancy rate in care homes by type of provision	S14	2007-2017	2
15. Places in care homes with en-suite facilities	S15	2007-2017	2	16. Body mass index (BMI) distribution of primary 1 education children	S16	2001-2019	1
17. Smoking prevalence among 13 and 15 year olds in Scotland	S17	2001-2019	2	18. Delayed discharges: monthly census	S18	2016-2020	2
19. % of children classed healthy weight, overweight, obese,severely obese at Primary 1 review	S19	2002-2015	1	20. Alcohol-related admissions (stays) or discharges	S20	1981-2019	20
21. Drug use among 13 and 15 year olds in Scotland	S21	2002-2015	3	22. Health care clients	S22	2016-2019	2

^{a}: acronyms for services names, ^{b}: years of existing records, ^{c}: number of factors (TS) per service pack

in Fig. 3. In Fig. 3 (a) we see relationships between *'Percent of births in Low birth weights'* and *'Home Care Clients in Home Care Group'*, in Fig. 3 (b) between *'Mental health problems . gender(female)'* and *'general practitioners (GP) . value'*. A multi-linear model with 5 dependent H&Sc factors (each for each plot is shown in Fig. 3(c) and with target *'Single rooms . care home sector(owned mortgage)'*. The pattern for the X-axes {*'N'*,'/ others' } means that the 1^{st} independent factor is drawn sequentially from the set *'Delayed discharges (DD's) . monthly census . ratio'*, *'Care home clients . gender(Males)'*, *'Alcohol use among young people . age(13-15)'*, *'Alcohol use among young people . age (15-18)'* while the rest of the independent's are the left *'N'*'s.

In ARMA models as we add more H&Sc factors more linear combinations are good but after adding more (>3 or 4) the number of successful ones (high linearity confidence) drops (as seen in Fig. 2 (a)). Fig. 2 (b) describes LR prediction where there is no drop. In Fig. 2 (c) we see the combined LR-CM prediction that is LR applied to same cluster's data. In Fig. 2 (c) the clusters are very close to the data or coincide with them when data are few. In LR-CM prediction we have normalized H&Sc factor (S22) and *le.r*= 0.8. In Fig. 2 (f) one can see CC based LR prediction for period (2004:2016) on H&Sc factor: S8. LR orders are from 2 to 6 (original is 7^{th} plot) applied to highly cross-correlated year series. The colors represent LR orders. The linear prediction for normalized H&Sc factor: (S2) in years (1981:1991) with variable orders (original, 1^{st} and 7^{th} shown) are shown in Fig. 2 (d).

Table III shows representative results for prediction using LR ($ts_1 = LR0 + LR1 * ts_2 + LR2 * ts_3 + ...$)) and ARMA ($ts = AR0 + AR1 * TS(t-1) + AR2 * TS(t-2) + ...$) and LR probabilities $P1$,$P2$,$P3$ and (c) NN's (3 hidden nodes) (c.1) backpropagation (*'BACKPROP'*), (c.2) resilient backtracking (*'RPROP+'* with weight, (c.3) resilient backtracking without weight (*'RPROP-'*)

The data were heterogeneous (various formats for dates or other counts), with missing years, numerical (ages), dichotomous (presence or not of a demographics class or ages bands), categorical (classes or long text descriptions instead of numbers). For example, ages were found both as ranges as in '...ages 65+ ' or as single numbers. The gender was a numerical tag '1' for 'Male' and '2' for 'Female'. Other records were counts of patients (without more specifications) or percentages. A breakdown of the indicative attributes and their levels per attributeis shown in Table I. Representative shapes for the factors as shown in Figs. 1 (a)-(f). The data contained up to 6 attributes (settings) per service and each attribute has possible values called 'levels'. The services without settings had one attribute ('Value'). Some data take up to 20 levels (as in 'Hospital Admission Reasons').

To evaluate the prediction methods error metrics are used as in [12] that discusses forecasts of the workload in an Emergency Department's (ED) using the ARMA model. The present work compares predictions using MAE, MRE, 'Root Mean Squared Error' (RMSE), and Services and settings can also be compared by associating administrative and clinical data in pairs of co-occurrence using a contingency 'dashboard' as in the 'matrix' method used by NHSS and discussed in [13]. Works in [14] or [15] and [16] use simulation algorithms to test the sensitivity of the number of patients to clinical events ('early diagnosis', 'critical clinical outcomes', etc.) and from there compute HC system's response errors. Arrival models are also used to predict the demand and especially the discharges rates as discussed in [17]. Then the error is the difference from the actual rate.

To set up the analysis framework PCA analysis was used. PCA determined as most important the services in the pack (H&Sc data frame) (S2) ('Smoking prevalence and deprivation (SALSUS)') with 10 H&Sc factors explaining 56.8 percent of the data variance while from the pack 'Alcohol-related

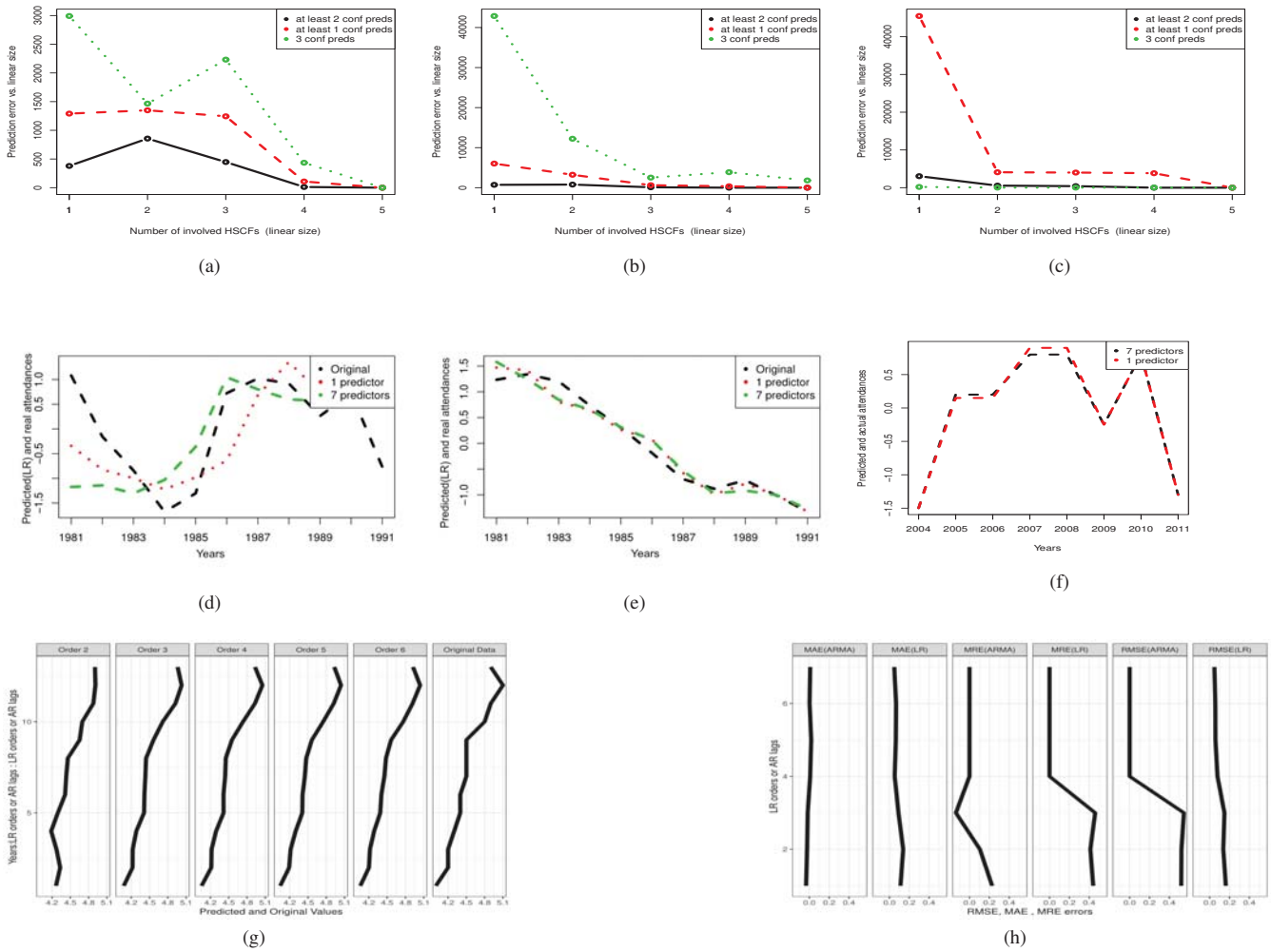


Fig. 2 Indicative observed attendances for LR, ARMA and their combination with CC and CM . The negative attendances are due to the normalisation and zero-padding

admissions (stays) or discharges (S20). 1 H&Sc factor explains 28.3 percent while for the rest of the packs most of the PCs are 2 or 3. Acronyms such as *'S.A.Z.'* are used to denote factor names that are part of service packs. The 1^{st} part, (S), of the acronym, is the ID of the service, then the acronym, *'A'*, of the attribute follows and then the ID of the level of the attribute follows as *'Z'*. For example the service *'Alcohol use among young people '* is (S1), the attribute for age *'A'* has levels *'Z''*s: {'13' ,'15', 'All'}. Each level was tracked as an individual factor or setting. This example indicates the number of patients aged 13 or 15 or any age 'All' tracked over the 39 years span. The services are also referred to in this work by their short names using Table II.

The main body of the statistical analysis included different types of linear regressions. Major prediction methods are reported in [10] such as Random Forests. In [18] Linear Regression (LR) is used to predict the work-load in hand surgery operations in the aging population or it is used to predict the clinical outcomes [19]. The roles of the different clinical factors as service attributes are analyzed in [20], [21] and in [12].

The ARMA model crashed for large lags (noted as p) or roughly $p > 4$. In [22] the ARMA models apply to likely linearly related year series. In [23] a step-wise regression is suggested as better over traditional LR or ARMA methods and this is also proven in this work empirically using specific ad-hoc ranges for LR orders and ARMA lags (delays). In [24] it is advocated that CC needs to be coupled with ML to reveal specific r elationships a cross d ata. T he L R m odels c an check for linear relationships among data determined by coefficients and probabilities as in [25]. In [26] LR is used to predict daily patient discharges using 20 patient features and 88 hospital ward level features and other administrative data. The basic formula used for LR is given in (1):

$$x_1 = \sum_{t=2}^{t=N} x_t \times a_t, t \in [2, N] \quad (1)$$

for some set of H&Sc observations (H&Sc factors), x_1, $x_2, ..., x_N$.

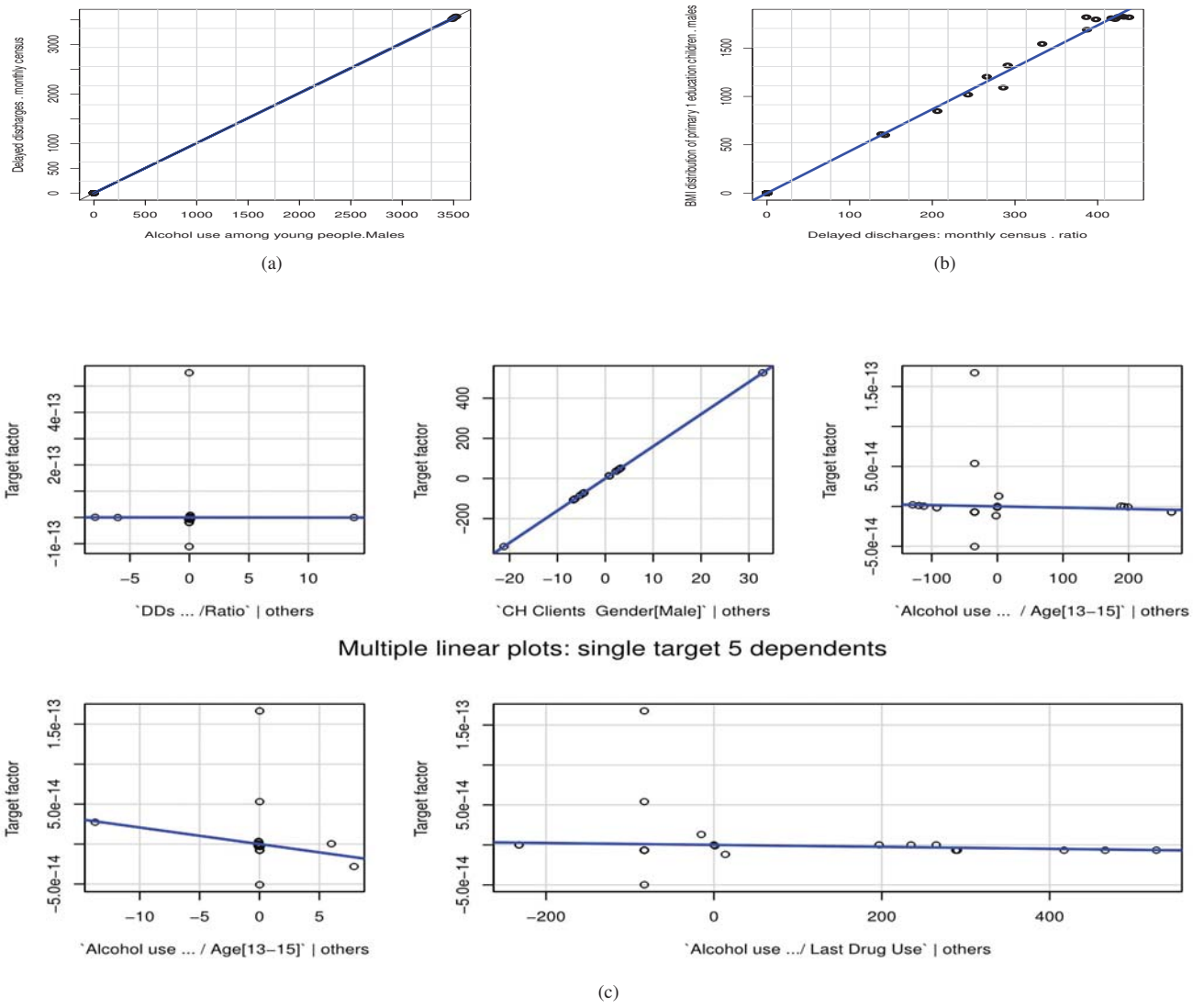


Fig. 3 Plots for indicative linear connections found ($HSc\ factor_i = LR(HScfactor_j), i <> j$). The captions below the individual plots show what are the independent variables

This work applied LR on NHSS data to relate different services attributes to client parameters to create cohorts so that similar data can be mutually predicted. The prediction error is given in (2):

$$R^2 = 1 - \frac{SSE}{SSR + SSE} = \quad 1 - \frac{\sum_{i=1}^{i=39} \left(\hat{y}_{39} - y_i\right)^2}{\left(\hat{y}_{39} - \hat{y_i}\right)^2 + \left(\hat{y}_{39} - y_i\right)^2} \tag{2}$$

where y_i stands for the LR representation of the i^{th}, $i \in [1, 39]$ year's attendance for some factor, while $\hat{y}_{39}$ stands for the average attendance across the 39 years. The prediction for year i and factor j_0 is $HSCF_{j_0} = \sum_{k=1}^{k \in [1, N_{group_{j_0}}] - [j_0]} a_k * HSCF_k$. The parameter $N_{group_{j_0}}$ is the number of all H&Scs in a linear relationship that predicts $HSCF_{J_0}$.

The classification of the servicesentailed prediction either by using LR or by using ARMA. For ARMA models time lags up to 2 or 3 were tried due to ARMA 's sensitivity to higher lags. ARMA predicts an H&Sc factor from its time-lagged samples. Comparisons were carried out between (a) ARMA, (b) CM-LR (CM using LR), (c) CC-LR (CC using LR), (d) LR prediction methods. Here the ARMA model used is as (3):

$$y(t)_{AR\ predicted} = \hat{y(t)} + \sum_{i=1}^{i=p} a_t * y(t-i) \tag{3}$$

for different orders, p.

The neural networks (NNs) worked well as predictors and were to predict service demands. The simplest NN for H&Sc data is a 3-layer network that has an input layer (I) that is a set of input data processing nodes $\vec{i} = [i_1, ..., i_{39}]^T$ where H&Sc data $hsc_i = [hsc_{i,1}, ..., hsc_{N_H,i}]^T$ come in, a 2^{nd} layer, (H) that is hidden, and is comprised from a variable number of processing nodes that receive weighted data sums

from (I), then scale, and threshold them (using an activation function), $F_j(\vec{hsc}) = F_j(\vec{hsc}^T * \vec{w_H}) = h_j$ for each node j in (H). N_H is the number of nodes at layer (H). Then (H) passes them to a third layer(Y), which is a variable set of output processing nodes that map the data to their final labels $y_k = NN(\vec{hsc_i}) = \vec{h_i}^T * \vec{w_Y}$ for input $\vec{hsc_i}$ during training or producing a prediction when new data come in. The (H) and (Y) layers are interlinked using a 2^{nd} set of weights $w_{HtowardY} = w_Y$. For prediction $N_Y = N_i = 39$. Finally, $F(x) = \frac{1}{1+exp^{-x}}$ is the cost function (the non-linearity). This is a 3-step process: $I \rightarrow H \rightarrow Y$. Unlike the regression methods seen so far the NNs are characterized by non-linearity and parallel processing. This allows NNs to better explore data inner correlations. The NN's weights were trained using 3 training algorithms (1) back-propagation ('BPROP'), (2) resilient backpropagation with weights ('RPROP+'), (3) resilient backpropagation without weights ('RPROP-'). As explained in [27] these differ in their weights convergence speed and in their weights updating algorithms but all are based on feeding back the prediction error to reach optimal weights. Indicative results for the 3 NNs are given in Table II under the column named 'NNs'. We can see that the NNs have a remarkably steady performance (RMSE=Er1) considering LR or ARMA. This is because the NNs have a more complex structure and convergence process than the LR or ARMA methods have which allows them to model the data better. On top of that, the NNs need to have normalized H&Sc data as in (4):

$$\hat{hsc} = \frac{hsc - max(hsc)}{max(hsc) - min(hsc)} \in [0, 1] \quad (4)$$

so that the cost function (f(x)) takes values $f(x) \in [0, 1]$ and data do not cause scale problems to the network.

The performance on the NNs on H&Sc data was also considered. As discussed in [28] the back propagation algorithms find a wide range of applications in HC operations. The above paper discusses using BPROP to predict the scenario where the length of stay (LoS) (hospitalization time) exceeds the average stay and learns from a training data set. LoS is a HC parameter that when predicted well using the correct operational parameters can save up to 2 days of stay. In this case, the performance is measured by the correct over-lengths (above average LoS). The NNs achieved in our work different RMSE errors with different layers. Also, the ROC analysis was used in this work that relies on the number of correct predictions of the i^{th} input in the outcomes $y_{i,k} = hsc_{i,k}$. Their binary counterparts for both $\vec{hsc}_i$ data and their predictions Y_i were calculated by rounding the actual values after maxmin normalizing them as in (4). Then a series of 1's and 0's was passed to ROC analysis for both of them as Fig. 4 shows.

III. Results and Discussion

Among the major findings of this work it was found that on average a number of factors in the region ([2,6]) were well linearly connected. For example 3- 5 independent factors and 1 dependent *'S2.Age.13'* was found as in Table II (3d row). Similar sizes and in the region ([2,5]) were also discussed in [29].There were no H&Sc factors that could not be expressed through linear combinations except for those with a single or a few (2) years records like *'Smoking prevalence among 13 and 15-year-olds in Scotland. health (Fair)'* (year:2017) or others with no records after 1997 or those with a single low attendance before 1997. Most H&Sc factors did not have records then. Also, those H&Sc factors with only very recent records, i.e., after 2017 and not before like *'Delayed discharges: monthly census. other living conditions*$< all\ levels >$*'* did not relate well (few cases).

Linear group members would be included (considered as well linearly linked to the same target) if their LR coefficients had low probabilities (low p-value for non-dependence) and the accuracy (RMSE,) was kept to an acceptable level ($>= 0.8$). The probability levels depended on the number of independent H&Sc factors used. The average observed was close to 1 percent and above 0.8. The accepted probabilities belonged to the interval ([0.001,0.05]). Extreme cases as below 0,001 or above 0,05, or were an over-fit or a non-fit and were ignored. of 1 was accepted but not with too low or too high probabilities. For LR the number of predictors was taken while for ARMA the number of lags (past samples). For LR and ARMA models the probabilities and the coefficients are computed (and referred to) for each prediction and a single MRE and MAE and RMSE error was used for the target H&Sc factor. The most often observed factors in various linear sets (as predictors) are: *'Alcohol Admissions'* (S22) (overall, i.e., summed over attributes and levels counts). The factor *'Alcohol-related admissions (stays) or discharges.care home Sector. voluntary'* (S20) (1981-2019) is the target in a combination that had 3 strong coefficients (predictors) with good probabilities {0.013,0,04} as seen in Table II (row 5) and had a low one (P(nonlinear coefficient) = 0.945). The pack *'Smoking prevalence and deprivation (SALSUS)'* (S2) is also the target in several linear combinations (rows #3 and #4 are in Table II and are only indicative). It is also interesting to observe how well the numbers of GPs per age band correlate. This can be very helpful for the planning of resources(that is the GPs). Such a combination has predictors (S12.A1.L1) *'Number GP registered patients . Age .16-64'*, (S12.A1.L1) *'Number GP registered patients . Age . All'*, with probabilities {P2=0.025, P3=0.046}. This is also confident because 2/3 of the predictors are in the crucial interval ([0.001,0.05]). Another interesting linear set is in row 4 which has 5 predictors and is a better linear approximation for the same pack's (S2) target (S2.A5.L1) *'Smoking prevalence and deprivation (SALSUS) . SIMD quintiles . 1 - most deprived'*. as in row #3 (S2.A1.L1) *'Smoking prevalence and deprivation (SALSUS) . Age . 13'* . The less populated linear group in row 3 has a stronger (lower nonlinearity ones) probabilities {0.002, 0.026} with respect to row #4 {0.096, 0.213}, and fewer factors (2) than the one in row 4 (has 5). The errors are comparable (in row 3 is 0.334 and in row 4 it is 0.483). One can find more combinations with a strong dependent factor and as many as 4 good independent factors in linear sets of size 5. The factors in the pack (S2) were common as dependent variables and were connected to several other factors as also discussed in [30] where the 30-day and 48-hour re-admission

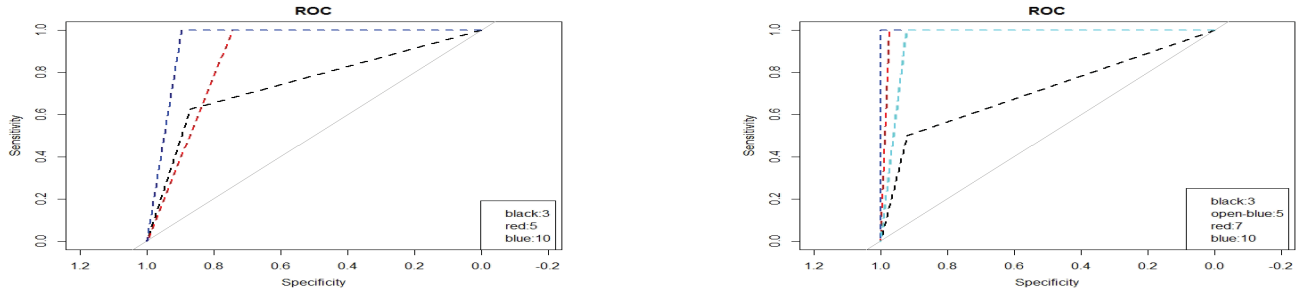


Fig. 4 ROC curves for 4 types of the *'RPROP'* with (H) layers

risks are computed using 7 reasons/factors that were not in the PHS data processed. The referred work uses an ARMA method considering re-admission drivers such as the number of re-admissions in the past 12 months. In [31] and [32] the number of the factors is actually a parameter to adjust which is in our case fixed, i .e., 1 10. I t w as f ound t hat very good independent factors the strongest coefficient belonged to the packs: (S20) *'Alcohol-related admissions (stays) or discharges'* and (S3) *'Smoking behavior and self-rated health (SALSUS)'* with many likely linear dependencies (that is, below the p.value of 0.05). Indeed, this can be expected as such causes are dominant in hospital admissions and are at the root of social problems. The probabilities that connect them to different factors are listed in Table II in columns labeled as *'P1'* and so on and for *'P1'* are 0.016 (target is (S1.A1.L3) *'Smoking prevalence in young people (SALSUS) . Age . All'*), or, 0.016 (target is: (S3.A4.L3) *'Smoking behaviour and self rated health (SALSUS) . Self-assessed general health . Good'*) as in the fifth and sixth rows in Table II. Also, a well-matching pack whose factors are used often as independent predictors is (S10) *'Mental wellbeing (SSCQ)'* as in the rows (2,3,4). This can be so because the mental problems (pack (S15)) cannot be isolated from smoking (packs (S2) and (S10), etc.) and might be related to a range of alcohol-relevant public services or patients cohorts who receive them. One of the factors was *'Percent of people aged 65+ who are admitted as an emergency to hospitals at least twice within 12 months'* that alone has connection probabilities $\{0.026, 0.035\}$ respectively to its predictors (*'Number of general practices with registered patients'*) and (*'Body mass index distribution of Primary 1 Education Children'*) that is not listed in Table II. In row 8 one can see that pack *'Mental wellbeing (SSCQ)'* (S15) is linked to distance health (pack (S14)*'Home intensive'*) that reveals the relationship between remote healthcare and mental problems. Zero padding revealed more relationships and did not limit the results only to common years. For example, the packs: (S1) (1998-2010) and (S3) (2008-2019) had very low overlap and although brought into the same span after zero-padding they were not found well linearly correlated as a pair but they were with other services. An example is *'Headcount of general practice workforce'* (S15) with *'Living arrangements for home care clients'* (2007-2017) (S14) while *'Alcohol use among young people'* is well connected with many but not with (S3). The service pack (S3) and especially its factor *'.type of tenure . owned loan'* is a well-modeled (predicted) factor and creates (where it is common) patients' categories as it can be seen in the same table with probabilities : $\{$1e-23,0.999,0.968$\}$ not in Table II). Some services connected with a p-value below 0.001 are likely an over-fit as in rows (6,8,9,12,13) in the same table.

An innovation of this work is the combination of the classification with regression schemes. The LR and ARMA models were used for training and testing data segments in an analogy (learning ratio learning ratiothat varied in the interval (I = [0.1,0.9]). The usual learning ratio is in the region ([0.6, 0.9]) as advocated in [33] where the learning cases are discussed (i.e., small, large learning ratios) with respect to the quality of the prediction. Here, lower learning ratios were used in the region ([0.2, 0.8]) due to the smoothness of the data that allowed easy learning at a low learning ratio. Some indicative results for LR are shown for sample dates (2004:2016) in Figs. 2 (a)-(h). These results were coupled with analysis using CM and CC as in Figs. 2 (f),(h),(e).

The PCA analysis defined a feature space in the wider H&Sc data space. The feature space was determined by the most important services that were found in the pack (H&Sc data frame) (S2) *'Smoking prevalence and deprivation (SALSUS)'* with 10 H&Sc factors explaining 56.8 percent of the data variance while from the pack *'Alcohol-related admissions (stays) or discharges'* (S20). 1 H&Sc factor explains 28.3% while for the rest of the packs most of the PCs are 2 or 3. Acronyms such as *'S.A.Z'* are used to denote the factors' names that are part of the services packs. The 1^{st}part, (S), is the ID of the service, then, A, is the acronym for attribute and then the ID of the level of the attribute is Z. For example the service *'Alcohol use among young people'* is (S1) and the attribute for age *'A'* has levels *'Z's'*: {'13 ' ,'15 ', 'All'}. Each level was tracked as an individual factor or setting. This example indicates the number of patients aged 13 or 15 or of any age *'All'* tracked over the 39 years span.

The AR prediction errors can be shown in Figs. 3 (a) and (b) illustrate how H&Sc factors are linearly related. Multiple linear plots with more than 2 independent H&Sc factors are shown in Fig. 2 (c). The RMSE was mainly used for prediction errors. The performance of the AR models did not always increase with increased lags as is the case with CC (predict

from the most cross-correlated) or with the LR models (predict using the more dependent). As the past samples increased the error could also remain stable as shown in Figs. 2 (a)-(c). The number of the samples of the model, (p), is a model's parameter. LR was tested on zero-padded data and an example is given for the factor *'Self-declared (SALSUS) smoking prevalence and deprivation. age. 15'* (or S2. A1. 15). Prediction results are shown in Fig. 2.

The RMSE was used for prediction errors. The RMSE for AR models did not always increase lags as was the case with CC (predict from the most cross-correlated) or with the LR (predict using the more dependent). By increasing past samples the error could remain stable as shown in Figs. 2 (a)-(c). The number of past samples, p, was a model's parameter. An example of using LR on zero-padded data is given for the factor *'Self-declared (SALSUS) smoking prevalence and deprivation . age. 15'* (or (S2. A1. 15)) (A1 is *'Age'*=15). . Prediction results are shown in Fig. 2.

As seen before, the PCA analysis could reduce the data dimensions to a feature space for the classification of the factors so that one can apply LR or ARMA in a lower dimensional space. PCA was applied on the 110 factors and gave 11 major eigen directions. Alcohol-related factors major PCs are they are more populated (20 attributes). The reasons for admission due to alcohol are more frequent, thus dominant due to their more likely variance. LR is facilitated when the services are represented in a PC's sub-space. For example services (S10.A2.L1) (*'Mental wellbeing (SSCQ) . Gender . All'*) and (S12.A6.L4) (*'Home care client living arrangements. Alone'*) predict (S2)'s services in rows #3 and #4 in Table II using the same data packs (as in row #3). The relationship between PCA and LR is beyond the scope of the present work. Other factors span more clusters like (S9) (*'Number of general practices. registered patients.*$< anylevel >$*'*) that concerns patients and services that are linked to GPs which can be more diversified since GP visits can be for different reasons.

When comparing PCA and LR we know that PCA is known for the minimal data representation it offers while LR is a way to model data using other data as their predictors [34]. The two approaches were compared as shown in Table II. PCA suggests, as in Table III, that the services pack 'S20' has all the PCs that is also confirmed in Table III where one can see that in many linear combinations the pack (S20) is a popular service (either dependent or independent) indicating that LR is in-line PCA. The scale difference of the RMSE is a matter of how many other variables are used.

For the prediction different learning/training ratios were used. CM defined the closest clusters that were used to train and test the LR or ARMA models and CC defined well cross-correlated limited data sets to train the models. LR-CM means CM followed by LR. The RMSE, MAE, MRE errors using CM and LR (LR orders up to 7) and AR (ARMA lags up to 3) are shown in Fig. 2. The LR-CM approach can be contrasted to cluster-wise regression discussed in [35] where the LR coefficients (LR structures) are the meta-data to cluster themselves.

In Fig. 4 we see the ROC curves of services tracking results using NN's. More specifically, in (a) the results using NN's with 3,5,7,10 nodes for H&Sc factor (ID=92) and triplet {S=12,A=13, L=1} which is *'Number GP . Registered Patients . Age . All'*, learning ratio=70%. As we add (H) layers the curves move towards the top left corner that is the perfect prediction case, and in (b) for the triplet {S=20,A=11, L=4} which is *'Home care.services . Value'*.

To benchmark the NNs ROC analysis was used and the results are shown in Fig. 4 which are obtained using *'BPROP'* on selected factors in Table III. The learning algorithms for all 3 algorithms used depended mainly on (H) nodes and less on learning ratio.

IV. Conclusion

The paper discussed how we can relate H&Sc services attendances using prediction to form services cohorts and evaluated several methods. The dependence of these relationships on classification as well as on services settings was studied using LR, AR and 3 types of NNs that used back propagation to learn. PCA and CM provided basic knowledge as to how can we limit the closest domain space for prediction. The results revealed that linearity holds for up to ca. 4 services and that LR works better than ARMA in regard to the accuracy of the prediction. Unlike common sense, NNs were less dependent on how well we trained them and more on the structure of them (hidden layer) as was revealed using ROC analysis and RMSE errors. Common years were more suited for linear relationships. LR methods worked better on low dimensions (few or selected years). AR models proved less successful with respect to LR as it is seen in the high RMSE, MAE, and MRE errors obtained. First groupings were found based on CC or CM were further explored using LR and ARMA that changed in the years. PCA yielded 11 best H&Sc factors and CM defined 5 main classes across the 39 years. The LR methods proved services are uncertain and may depend on factors such as the year the data were recorded [15]. Some H&Sc factors were found to be widely attended such as the Emergency Departments works and highly cross-correlated to less attended H&Sc factors. The work revealed that services that are more common as predictors with other services were related to *'Alcohol Admissions'* as for example (S20) and home-based (various services : (S11), (S12), (S14), etc.) services and confirmed these are common reasons for getting admitted to a hospital and that services may expand and differentiate once a patient is originally admitted for one of these reasons. Moreover, the HC system has grown around services offered to the elderly or to home-based users as seen by the plethora of services offered from a distance and their participation to more services groupings. The high specialization of services offered to alcohol-related patients was confirmed by the high linear confidence attached to such H&Sc factors as low births weights and services related to alcohol. Depending on the year at hand, though, the *'....low birth weight (weight* $< 2500gr$*) '* class can also be regressed (linearly related) with mental health patients [36]. It was also found that GPs workforce could be related to patients self-assessed as being well (SALSUS). Among other findings, low

birth weights are related to the people who are offered housing on a voluntary basis in care homes and both are linearly related to the patients that are registered with GPs and live in adult-type care homes. These may offer links across the data that may not be expected or even justified. The merits of using ML is that it can offer out off the box solutions that may offer insights as for hidden data relationships.

TABLE III
INDICATIVE LINEAR GROUPS WITH LR MODEL PARAMETERS

Linear groups of H&Sc factors[1] IDs [2]									
LR				ARMA			NNs		
Er1	LR0 LR1 LR2	P0 P1 P2	Er2 Er3	Er1	AR0 AR1 AR2	Er2 Er3	NNs(3 layers[*5])	NNs(10 layers)	NNs(15 layers)
[1] **(1)** S20.A9.L5 [*4], **(2)**S3.A3.L3, **(3)** S20.A3.L2									
0.167	0.016	0.993	0	8.852	-27e-4	0	0.372	0.159	0.159
	0.006	0.013	1		-27e-4	1	0.406	0.022	0.029
	-2e-4	0.040			8e-16		0.371	0.02	0.166
[2] **(1)**S20.Care Home Sector.Voluntary [*3], **(2)** S10.A3.L1, **(3)** S12.A8.L1									
0.945	25264	0.04	0	7609	0.93	0	0.32	0.343	0.0429
	40.6	0.001	1		-73e-4	1	0.322	0.08	0.385
	184	0.049			37797		0.322	0.059	0.039
[3] **(1)** S2.Value, **(2)** S10.A8.L1, **(3)** S12.A6.L4									
0.334	-13.6	0.366	0	95.5	-27e-3	1	0.321	0.0216	0.0259
	0.042	0.002	1		-27e-3	0	0.323	0.026	0.029
	0.263	0.026			15.6		0.321	0.027	0.029
[4] **(1)** S2.Value, **(2)** S10.A8.L1, **(3)** S12.A6.L4, **(4)** S3.A3.L1, **(5)** S20.A9.L2, **(6)** S10. Limiting Physical/Mental Condition.Limiting condition [*4]									
0.483	2.361	0.024	0.025	7.605	-0.3	230	0.321	0.141	0.026
	0.0173	0.096	1		0.99	4	0.316	0.159	0.158
	4e-3	0.213			0.195		0.316	0.028	0.148
[5] **(1)** S3.A7.L3, **(2)** S20.A3.L1, **(3)** S4. Client group in care home.Adults with disabilities [*1]									
0.821	24	0.026	1	0.98	-0.08	0	0.406	0.224	0.255
	0	0.035	119.5		70	1	0.397	0.34	0.313
	1e-3	0			63		0.397	0.361	0.358
[6] **(1)** S20 . A3. L1, **(2)** S20 . A3. L2, **(3)** S12 . A6. L1									
0.237	-15	0.521	3e-15	142	-27e-3	0 1	0.396	0.158	0.159
	0.003	0.025	1		-27e-3		0.398	0.158	0.159
	0.423	0.024			23		0.398	0.157	0.139
[7] **(1)**S20 . A7.L3, **(2)** S10.A8.L1, **(3)** S13.A1.L3									
0.066	26445	0.002	2	4.366	0.89	22380	0.35	0.214	0.183
	-7.62	0.331	1		-15e-2	0	0.351	0.217	0.243
	0.973	0.247			59e-3		0.356	0.217	0.243
[8] **(1)** S12.A8.L3, **(2)** S12.A8.L4, **(3)** S20.A5.L2									
0.049	1.654	5.964	0.608	2.1e-8	0.962	10309	0.391	0.039	0.043
	-6.53	0.25	1		-0.061	0	0.403	0.097	0.104
	-0.002	0.457			-0.041		0.4	0.251	0.395
[9] **(1)** S12.A8.L3, **(2)** S20.A3.L1, **(3)** S4.A7.L3									
0.389	0.923	0.005	0.6	2.1e-8	0.962	1.193	0.391	0.0184	0.027
	-1e-4	0.604	1		-0.061	1	0.403	0.025	0.158
	5e-4	35e-6			-0.041		0.4	0.025	0.16
[10] **(1)**S20 . Value, **(2)** S10 . A8.L3, **(3)** S20 . A7. L4									
0.013	34.592	0.24	3e-9	930	0.705	271	0.32	0.322	0.321
	-7.37	0.604	1		-96e-5	1	0.343	0.039	0.0215
	-3e-4	0.977			-0.261		0.043	0.059	0.029
[11] **(1)**S3 . A3 . L1, **(2)** S20 . A2. L1, **(3)** S17 . A5. L3									
0.197	3e-3	0.009	1 7.604	0.989	0.195	3.9	0.183	0.214	0.366
	0.0006	0.944			1e-23	0	0.243	0.217	0.351
	0.023	0.25			-0.32		0.243	0.214	0.35
[12] **(1)** S3 . A6 . L2, **(2)** S11 . A10 . L1, **(3)** S17. A5. L3, **(4)** S17 . A6 . L2									
0.349	0.08	3e-15	1e-23	929.98	0.705	0	0.391	0.251	0.395
	6.89	0.999	1		-96e-5	271	0.403	0.097	0.102
	-0.005	0.968			-0.261		0.4	0.251	0.046
[13] **(1)** S5 . A3. L1, **(2)** S3. A7 . L3 , **(3)** S3 . A7. L2									
0.998	24.2	0.615	0.45	2e-14	1	-0.03	0.255	0.297	0.398
	0.406	0.675	0.83		-0.03	0	0.234	0.361	0.158
	-0.139	1e-23			-0.03		0.406	0.358	0.159

[*1] H&Sc factorsare shown as triplets $x.y.z$ (c.f. section II), [*2] attributesand levels as per Table-2, 1^{st} and 2^{nd} columns, [*3] naming convention not shown(but used), [*4] are the levels of attributes(some are listed in Table I) [*5] are the numberof nodes in (H) layer

ETHICAL APPROVAL

There are is no animal experimentation in the manuscript and no needed to be obtained. No studies on humans were carried out.

DATA AVAILABILITY

The data used in this work were made freely available online by PHS as an open database. The link is provided in the references.

FUNDING

The author received no specific financial support for the authorship and/or publication of this article. The author during the works of this paper was funded by an Abertay University, Dundee stipend.

AUTHORS CONTRIBUTIONS

The author is the only writer of the manuscript and carried out all the research. Other help or contributions are acknowledged as well as the data used.

ACKNOWLEDGMENT

The author is grateful to PHS and to NHSS and especially to Mr. Martin McKenna, Mr. Andrew Mooney, Dr. Lee and Mr. Paul Leak (Scottish Government) for accommodating in their technical meetings, for sharing ideas, bibliographies that helped and guided this work. The author is also thankful to Dr. James Bown, Dr. Euan Dempster and Dr. Ann Savage from Abertay University Dundee for comments in early phases of this work. The author during the works of this paper was funded by an Abertay University Dundee stipend.

REFERENCES

[1] Simon Bottery . $https : //www.kingsfund.org.uk/about - us/whos - who/simon - bottery?page = 2$. The King's Fund
[2] Public Health Scotland (2020). Data and intelligence. A – Z Subject Index. $https : //www.isdscotland.org/A - to - Z - index/index.asp$
[3] Scottish Government (2019). Statistics Service Health and Social Care Data. $https : //statistics.gov.scot/datahome$
[4] Vittorio Lippi (2019). Incremental Principal Component Analysis: Exact implementation and continuity corrections. arXiv: 1901:07922v2; stat:ML; 13May2019. $https : //arxiv.org/pdf/1901.07922.pdf$
[5] Ian Litchfield (2019). Can pathways of patients with long-term conditions in UK primary care? A study protocol. BMJ Open, 2018. $https : //bmjopen.bmj.com/content/8/12/e019947$
[6] Dimitris Bertsimas (2018), Colin Pawlowski, Ying Daisy Zhuo (2018). From Predictive Methods to Missing Data Imputation: An Optimisation Approach. Journal of Machine
[7] E.M. Mirkes (2018), T.J. Coats, J. Levesley, A.N. Gorban (2018). From Predictive Methods to Missing Data Imputation: An Optimisation Approach. Journal of Machine Learning Research, 18 (2018),1-39. $http : //dx.doi.org/10.1016/j.compbiomed.2016.06.004$
[8] deRooij M. (2018). Transitional modeling of experimental longitudinal data with missing values. Adv Data AnalClassif, 12,107–130. $https : //link.springer.com/article/10.1007/s11634 - 015 - 0226 - 6$
[9] Burd M. (2015) . The Health System Matrix 6.1: Understanding the Health Care Needs of the British Columbia Population through Population Segmentation. $https : //doi.org/10.1186/s13643 - 019 - 1105 - 6$
[10] Gredell Devin (2019). Comparison of Machine Learning Algorithms for Predictive Modeling of Beef Attributes Using Rapid Evaporative Ionization Mass Spectrometry (REIMS). Data. Sci Rep.,9 5721 (2019). $https : //pubmed.ncbi.nlm.nih.gov/30952873/$
[11] Md Saiful Islam, Md Mahmudul Hasan (2018). A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining, Healthcare, 2018,6-54. doi : $10.3390/healthcare6020054$
[12] Muge Capan (2020), Stephen Hoover, et al. (2019). Time Series Analysis for Forecasting Hospital Census: Application to the Neonatal Intensive Care Unit Multitask learning and benchmarking with clinical time series data. Appl. Clin. Inform.. 2019,7(2):275–289. $https : //dx.doi.org/10.4338\%2FACI - 2015 - 09 - RA - 0127$
[13] Langton, J.M. (2018), Wong, S.T., Burge, F. et al. (2015). Population segments as a tool for health care performance reporting: an exploratory study in the Canadian province of British Columbia. BMC Fam Pract,21-98(2020). $https : //doi.org/10.1186/s12875 - 020 - 01141 - w$
[14] Vimal Mishra (2019), MD, MMCI, Shin-Ping Tu, MD, MPH, Joseph Heim, PhD, Heather Masters, MD, Lindsey Hall, MPH, Ralph R. Clark, MD, Alan W. Dow, MD (2019). Predicting the Future: Using Simulation Modeling to Forecast Patient Flow on General Medicine Units. J. Hosp. Med.,2019,1,9-15. doi:$10.12788/jhm.3081$ [10]
[15] Guersel, Gueney (2019). Healthcare, uncertainty, and fuzzy logic. Digital Medicine ,2016,2,101-12. $https : //www : researchgate : net = publication = 310817255Healthcareuncertaintyandfuzzylogic$
[16] Deborah A.Marshall, LinaBurgos-Liz et al. (2015). Applying Dynamic Simulation Modeling Methods in Health Care Delivery Research—The SIMULATE Checklist:Report of the ISPOR Simulation Modeling Emerging Good Practices Task Force. Value in Health, volume 18,Issue 2,March 2015,143-144. $https : //doi.org/10.1016/j.jval.2014.12.001$
[17] D. Ben-Tovim, J. Filar, et al. (2019). Hospital Event Simulation Model: Arrivals to Discharge. 21st International Congress on Modelling and Simulation. Gold Coast,Australia. $https : //www.mssanz.org.au/modsim2015/H2/bentovim.pdf$
[18] Bebbington E, Furniss, D . (2015). Linear regression analysis of Hospital Episode Statistics predicts a large increase in demand for elective hand surgery in England. J. Plast. Reconstr. Aesthet. Surg, 2015, Feb,68(2),243-51. doi:$10.1016/j.bjps.2014.10.011$
[19] Uematsu, H., Yamashita, K., Kunisawa, S., Otsubo, T., & Imanaka, Y. (2017). Prediction of pneumonia hospitalization in adults using health checkup data. PloS one,12(6),e0180159. $https : //doi.org/10.1371/journal.pone.0180159$
[20] Juang WC, Huang SJ, Huang FD, Cheng PW, Wann SR. (2017). Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. BMJ Open, 2017, Dec 1,7(11),e018628. DOI: $10.1136/bmjopen - 2017 - 018628$
[21] Harutyunyan, H., Khachatrian, H., Kale, D.C. et al. (2019). Multitask learning and benchmarking with clinical time series data. Sci .Data,6,96(2019).$https : //doi.org/10.1038/s41597 - 019 - 0103 - 9$
[22] Bui C., Pham N., Vo A., Tran A., Nguyen A., Le T. (2017). Time Series Forecasting for Healthcare Diagnosis and Prognostics with the Focus on Cardiovascular Diseases. Vo Van T.; Nguyen Le T.; $https : //link.springer.com/chapter/10.1007/978 - 981 - 10 - 4361 - 1138$
[23] Liew, B.X.W., Peolsson, A., Rugamer, D. et al. (2020). Clinical predictive modelling of post-surgical recovery in individuals with cervical radiculopathy: a machine learning approach. Sci.Rep,10,16782(2020). $https : //doi.org/10.1038/s41598 - 020 - 73740 - 7$
[24] Dunsmuir WT (2019). Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. Behav Res Methods,2016,Jun,48(2),783-802. DOI:$10.3758/s13428 - 015 - 0611 - 2$
[25] Skiera, Bernd & Reiner, Jochen (2018). Regression Analysis. Homburg, Christian, Klarmann, Martin, Vomberg, Andreas (Editors). Handbook of Market Research. $https : //doi : org = 10 : 1007 = 978 - 3 - 319 - 05542 - 8\_17 - 1$
[26] Shivapratap Gopakumar (2016), Truyen Tran, Wei Luo, Dinh Phung, . JMIR Medical Informatics 4(3):e25 . DOI: $10.2196/medinform.5650$.
[27] Ciprian Florescu & Christian Igel (2018), RESILIENT BACKPROPAGATION (RPROP) FOR BATCHLEARNING IN TENSORFLOW. Work-

shop track - ICLR 2018 . $https$: $//openreview.net/pdf?id$ = $r1R0o7yDz$

[28] Ippoliti, R., Falavigna, G., Zanelli, C. et al. . Neural networks and hospital length of stay: an application to support healthcare management with national benchmarks and thresholds. . Cost Eff Resour Alloc 19, 67 (2021). $https$: $//doi.org/10.1186/s12962 - 021 - 00322 - 3$

[29] Yang, C., Delcher, C., Shenkman, E. et al. (2019). Expenditure variations analysis using residuals for identifying high health care utilizers in a state Medicaid program. BMC Med Inform Decis Mak, 19,131(2019). $https$: $//doi.org/10.1186/s12911/019/0870/4$

[30] Daniel J. Morgan, Bill Bame, Paul Zimand, et al. (2019). Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. JAMA Netw Open,2019,Mar,2

[31] Boelaert, Julien & Ollion, Etienne. (2018). The Great Regression. Machine Learning, Econometrics, and the Future of Quantitative Social Sciences. Revue française de sociologie. 59.

[32] Marno Verbeek . A Guide to Modern Econometrics . John Wiley & Sons. DOI10.3917/rfs.593.0475. $https$: $//www.researchgate.net/publication/$227488993 $_A_Guide_to_Modern_Econometrics$

[33] Aitor Lewkowycz and Ethan S Dyer and Guy Gur-Ari and Jascha Sohl-dickstein and Yasaman Bahri (2020) . ICLR 2021 Conference . The large learning rate phase of deep learning . $https$: $//arxiv.org/abs/$2003.02218

[34] Liu C, Zhang X, Nguyen TT, et al. (2021). Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches. General Psychiatry,2022; 35 : e100662(2021). doi: $10.1136/gpsych - 2021 - 100662$

[35] Torti, F., Perrotta, D., Riani, M. et al. Assessing trimming methodologies for clustering linear regression data. Adv Data Anal Classif 13, 227–257 (2019). https://doi.org/10.1007/s11634-018-0331-4

[36] Lyall DM, Inskip HM, Mackay D, Deary IJ, McIntosh AM, Hotopf M, Kendrick T, Pell JP, Smith DJ. Low birth weight and features of neuroticism and mood disorder in 83545 participants of the UK Biobank cohort . BJPsych Open. 2016 Jan 28,2(1):38-44. DOI:10.1192/ $bjpo.bp$.115.002154. $PMID$: 27703752, $PMCID$: PMC4995581