Original Research Article

Designing Social Care Plans by Grouping Services and Patients in Mixed Cohorts. A Study Using Regression Versus Neural Nets

ABSTRACT

Aims: Linking social needs to social classes using different criteria may lead to social services misuse. The paper discusses using ML and Neural Nentwoks(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 servicesters.
Study design: The work is based on public data from 22 services offered by Public Health Services (PHS) Scotland that break down into 110 years series called factors
Place and Duration of Study: NHSS and Abertay University, Dundee, from 2018 to 2020
Methodology: The paper discusses using ML and Neural Nentwoks(NNs) . The paper combines typical regression models with clustering and cross-correlation as complementary constituents to predict the demand. uses Linear Regression (LR), Autoregression (ARMA) and 3 types of backpropagation(BP) Neural Networks(BPNN) to link them under specific conditions.
Results: Relationships found were between smoking related healthcare provision, mental health related health serices, and epidemilogical weight in Primary 1(Education) Body Mass Index(BMI) in chlildren. Primary component analysis (PCA) found 11 significant factors while C-Means(CM) clustering gave 5 major factors clusters.
Conclusion: 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.

Keywords: Probability, cohorts, data frames, services, prediction

1. INTRODUCTION

There have been concerns about the system of publicly funded social care in England and in Scotland for more than 20 years (Simon Bottery, 2018). The present work advocates that additional revenue can be created to face the growing demand by connecting services using classification and prediction. That work states that it can be as high as £12 billion by 2030/31 at an average rate of 3.7 percent a year. The data were public H&Sc data available on PHS' website (Scottish Government, 2019) posted by June, 2019. The data used here were counts of patients (called 'Value' attribute in the data) and contained the parameters for each service. PCA was applied to see the most important ones after normalization was applied as discussed in (Vittorio Lippi, 2019).
Works that mine services sequences(patterns) belong to the same category as they use similarity metrics to stored patterns (Ian Litchfield, 2019) 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 (Dimitris Bertsimas, 2018) and for imputation using Markov models in (E.M. Mirkes, 2018), (de Rooij M, 2018) that use statistical models to approach the missing data.
Linear prediction as for example Auto-regressive Moving Average (ARMA) is discussed in (Langton, J.M., 2018) while the linear association of different services parameters is also the

question in (Gredell Devin, 2019) and a review of linear methods in the healthcare (HC) is presented in (Md Saiful Islam, 2018).
The paper is organized as follows. In the 1st section the nature of the data processed is better explained and the main analysis is given by introducing the LR and its application on the PHS data (Public Health Scotland, 2020). The ARMA / AR prediction is presented along with cross-correlation (CC), C-Means (CM), and PCA. Indicative comparisons and results are presented in the 2nd 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).

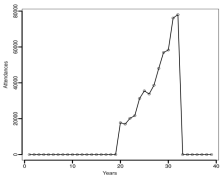
2. Materials and methods

2.1 The data

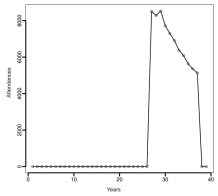
The data were heterogeneous (various formats for dates or for other counts), with missing years, numerical(ages), dichotomous (presence or not of a demographics class or ages bands), categorical (classes or long text descriptions instead of numbers). For example, ages were found both as ranges as in '...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 attribute is shown in Table 1.
Representative shapes for the factors as shown in figures 1(a)-1(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').

2.2 Error metrics

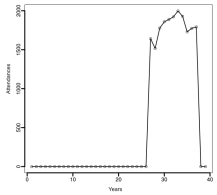
To evaluate the prediction methods metrics are used as in (Muge Capan, 2020) that discusses forecasts of the work-load in an Emergency Department's (ED) using the ARMA model. The present work compares predictions using MAE, MRE, 'Root Mean Squared Error'(RMSE), and Services and settings can also be compared 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 (Langton, J.M., 2018). Works in (Vimal Mishra, 2019) or (Guersel, Gueney, 2019) and (Deborah A.Marshall, 2015) 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 (D. Ben-Tovim, 2019). Then the error is the difference from the actual rate.



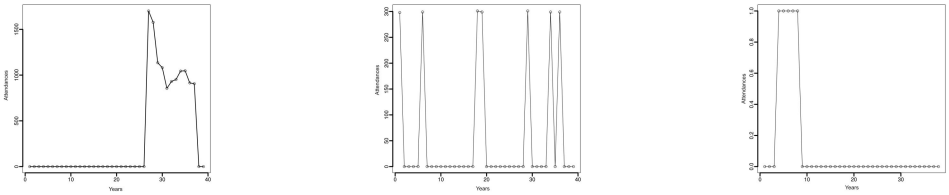
(a)



(b)



(c)



(d) (e) (f)

Fig.1 Most representative and varying H&Sc factors across the H&Scs groups: x-axis are the years and y-axis are the attendances per factor. (1a) S1 . A2.L2('*Smoking prevalence and deprivation(SALSUS) . Gender . F*'), (1b) S8.A8.L1('*Occupancy rate in care homes . Household type . Adults*'), (1c) S4. A1. L1('*Primary1 children BMI epidemiological . Age . 13*'), (1d) S7 . A8. L4('*Places in care homes . Household type . with children*'), (1e) S8.A8.L3('*Occupancy rate in care homes by type of provision . Household type . Pensioners*' , (1f) S16 .Value('*BMI distribution of primary 1 education children . Value*') . X-axis shows the span of 39 years , Y-axis shows the attendances , "L"s are the levels, some are shown in Table 1

2.3 Experimental Design

PCA analysis designated as the 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 admissions (stays) or discharges* '(S20). 1 H&Sc factor explains 28.3 percent while for the rest of packs most of the PCs are 2 or 3. Acronyms such as 'S.A.Z.' are used to denote factors names that are part of services packs. The 1st 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 2.

3. Research Method

3.1. Statistical Analysis

Linear regressions: Major prediction methods are reported in (Gredell Devin, 2019) such as Random Forests. In (Bebbington E, 2015) Linear Regression (LR) is used to predict the work-load in hand surgery operations in aging population or it is used to predict the clinical outcomes in (Uematsu, H, 2017). The roles of the different clinical factors as services attributes are analyzed in (Juang WC, 2017), (Harutyunyan H., 2019) and in (Muge Capan, 2020).

The ARMA model crashed for large lags(noted as) or roughly . In (Bui C., 2017) the ARMA models apply to likely linearly related year series. In (Liew, B.X.W., 2022) 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 (Dunsmuir WT, 2019) it is advocated that CC needs to be coupled with ML to reveal specific relationships across data. The LR models can check for linear relationships among data determined by coefficients and probabilities as in (Skiera, Bernd, 2018). In (Shivapratap Gopakumar, 2019) 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 below in equation 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), $x_{1,} x_{2,} .., x_{N.}$

3.2 Using LR on NHSS data

LR relates different services attributes to clients parameters to create cohorts so that similar data can be mutually predicted. The prediction error is given in equation 2:

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

, 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: $HSCF_{j_0} = \sum_{k=1}^{k \in [1, N_{group_{j_0}}] - [j_0]} a_k \square HSCF_k$.

The parameter $N_{group_{J_0}}$ is the number of all H&Scs in a linear relationship that predicts $HSCF_{J_0}$.

3.3 Using ARMA prediction with classification

For ARMA models time lags up to 2 or 3 were tried due to ARMA 's sensitivity to higher lags. ARMA predicts a 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 in equation 3:

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

, for different orders, *p*.

3.4 Neural Networks as predictors

The NNs can be trained to predict service demands. The simplest NN for HSc data is a 3-layer network that has an input layer(I) that is a set of input data processing nodes ($i = [i_{1,} .., i_{39}]$) where HSc data ($hsc_i^T = [hsc_{i,1}, ..., hsc_{N_H,i}]^T$) come in, a 2nd 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), $F_j(hsc) = F_j(hsc^T \square w_H) = h_j$, $w_{I \to H} = w_H$ 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 ($y_k = NN(hsc_i) = h_i^T \square w_Y$) ,

$h_{i,j}^T = [h_{1,j}, \ldots, h_{39,N_H}]^T$ for input hsc_i during training, or, produce a prediction when new data come in. The (H) and (Y) layers are interlinked using a 2nd set of weights ($w_{H \to Y} = w_Y$) . 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: $I \to H \to Y$.

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) backpropagation ('BPROP'), (2) resilient backpropagation with weights ('RPROP+'), (3) resilient backpropagation without weights ('RPROP-'). As explained in (Ciprian, 2018) 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 2 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 equation (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.

3.5 NNs performance on H&Sc data

As discussed in (IppolitiR, 2021) 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 $y_{i,k} = hsc_{i,k}$ s . Their binary counterparts for both hsc_i data and their predictions Y_i were calculated by rounding the actual values after maxmin normalizing them as in equation (4). Then a series of 1's and 0's was passed to ROC analysis for both of them. It was interesting to observe that the learning rate (*le.r*) or the 3 compared learning algorithms affected less the performance than the number of (H) layers did as figure (5) shows.

4. Results and Discussion

4.1 Major findings

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 2(3d row). Similar sizes and in the region ([2,5])were also discussed in (Yang, C.,

2019).

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 2(row 5) and had a low one (P(nonlinear coefficient)= 0.945). The pack 'S2'('*Smoking prevalence and deprivation(SALSUS)*') is also the target in several linear combinations(rows #3 and #4 are in Table-2 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 stronger(lower nonlinearity ones) probabilities ({0.002, 0.026}) with respect to row #4({0.096, 0.213}) and fewer factors (2) that 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 is also discussed in (Daniel J., 2019) where the 30-day and 48-hour re-admission 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 past 12 months' number of re-admissions.

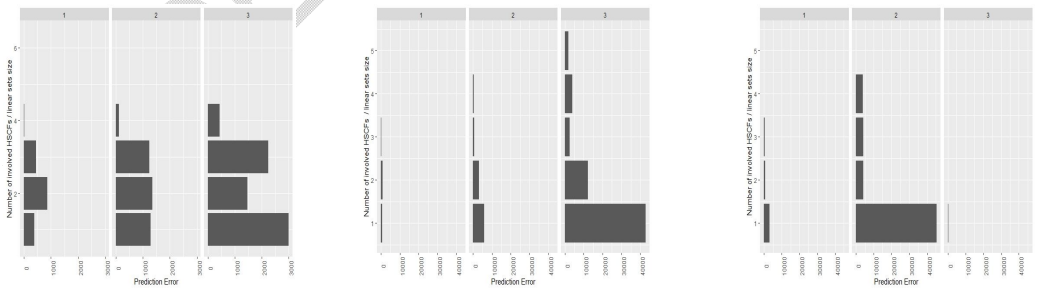
In (Julien Boelaert, 2018) and (Marno Verbeek, 2017) the number of the factors is actually a parameter to adjust which is in our case fixed, i.e., 110. It was found that vey good independent factors the strongest coefficient beloged do the packs: "S20"("*Alcohol-related admissions (stays) or discharges*")and "S3"("*Smoking behaviour 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 for hospital admissions and are at the root of socal problems. The probabilities that connect them to different factors are listed in Table 2 in columns labelled as 'Pr1' 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 2. 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","S10",etc) and may be related to a range 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 2. In row #8 one can see that pack "*Mental wellbeing(SSCQ)* "("S15") is linked to diatance health (pack "S14"("*Home intensive*")) that reveals the relationship between remote healhtcare and mental problems.

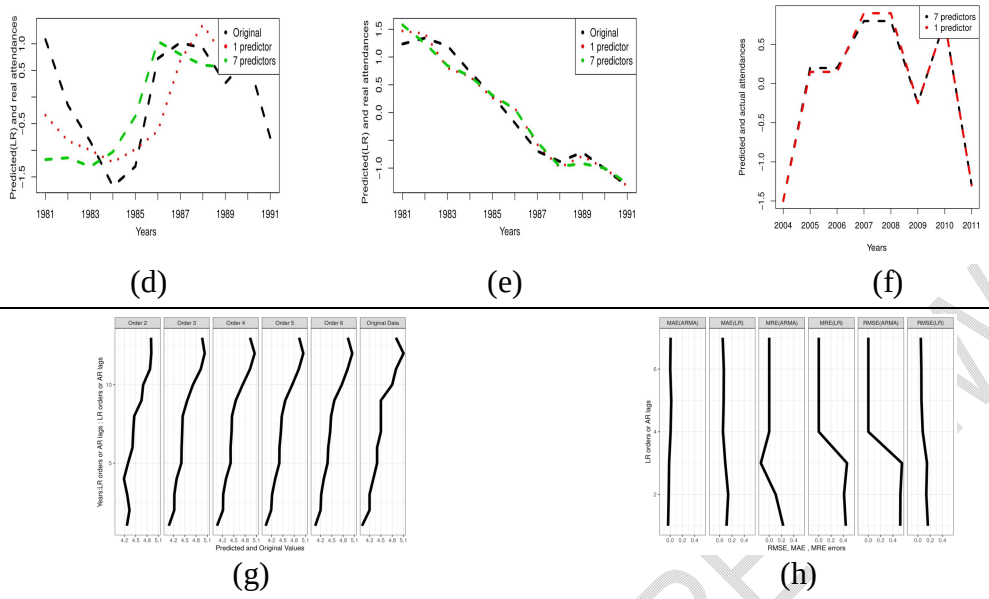
Zero padding revealed more relationships and did not limit the results only to common years. For example, the packs: 'S1'(1998-2010) and 'S3'(2008-2019) had very low overlap and although brought into the same span after zero-padding they were not found well linearly correlated as a pair but they were 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 modelled (predicted) factor and creates(where it is common) to patients categories as it can be seen in the same table with probabilities : ({1e-23,0.999,0.968} not in Table-2). 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.

4.2 Combining classification with regression schemes

The LR and ARMA models were used for training and testing data segments in an analogy(learning ratio (le.r)) that varied in the interval (I = [0.1,0.9]). The usual le.r is in the region ([0.6, 0.9]) as advocated in (Aitor Lewkowycz, 2020) where the learning cases are discussed (i.e., small, large le.rs) with respect to the quality of the prediction. Here, lower le.r's were used in the region ([0.2, 0.8]) due to the smoothness of the data that allowed easy learning at a low le.r. Some indicative results for LR are shown for sample dates (2004:2016) in figures 2(a) to 2(h). These results were coupled with analysis using CM and CC as in figures 2(f), 2(h), 2(e).



(a) (b) (c)



(d) (e) (f)

(g) (h)

Fig.2 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, (2a) ARMA, (2b) LR, (2c) LR-CM, (2d) Linear prediction for normalized H&Sc factor : 'S2' in years (1998:1991) with variable orders (original, 1st and 7th shown), (2e) LR-CM prediction with variable orders original, 1st and 7th normalized shown, and selected years (1981:1991) for 'S20'('*Alcohol-related admissions (stays) or discharges*") for *le.r*= 0.3,(2f) LR-CM prediction for normalized H&Sc factor 'S22'('*Health care clients*') and *le.r*= 0.8 (clusters are very close to the data or coincide with them when data are few),(2g) LR, ARMA errors compared for H&Sc factor 'S2.Value'('S*moking prevalence and deprivation*') using few AR lags(1 to 6) and LR orders and RMSE, Median Absolute Error and Median Relative Error errors. 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 CM that are not shown, (2h) CC based LR prediction for period (2004:2016) on H&Sc factor : S8('*Occupancy rate in care homes by type of provision*'). LR orders from 2 to 6(original is 7th plot) applied to highly cross-correlated year series. The colors represent LR orders. Negative attendances are due to normalisation and zero-padding .

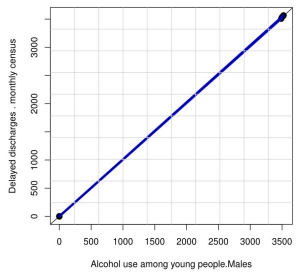
4.3 AR prediction errors

Figures 3(a), 3(b) illustrate how H&Sc factors are linearly related. Multiple linear plots with more than 2 independent H&Sc factors are shown in figure 3(c).

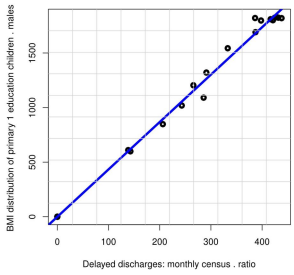
The RMSE was mainly used for prediction errors. The performance of the AR models did not always increase with increased lags as is the case with CC (predict from the most cross-correlated) or with the LR models (predict using the more dependent). As the past samples increased the error could also remain stable as shown in figures 2(a)-2(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 figures 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

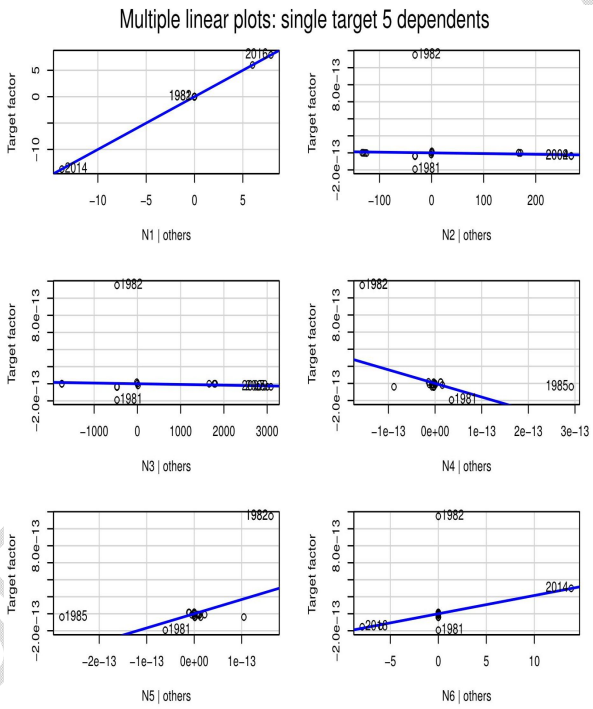
stable as shown in Figs. 2(a-c). The number of past samples, *p,* was a model's parameter. An example of using LR on zero padded data is given for the factor '*Self-declared(SALSUS) smoking prevalence and deprivation . age. 15'* (or 'S2. A1. 15')(A1 is 'Age'=15). Prediction results are shown in figures 2.



(a)



(b)

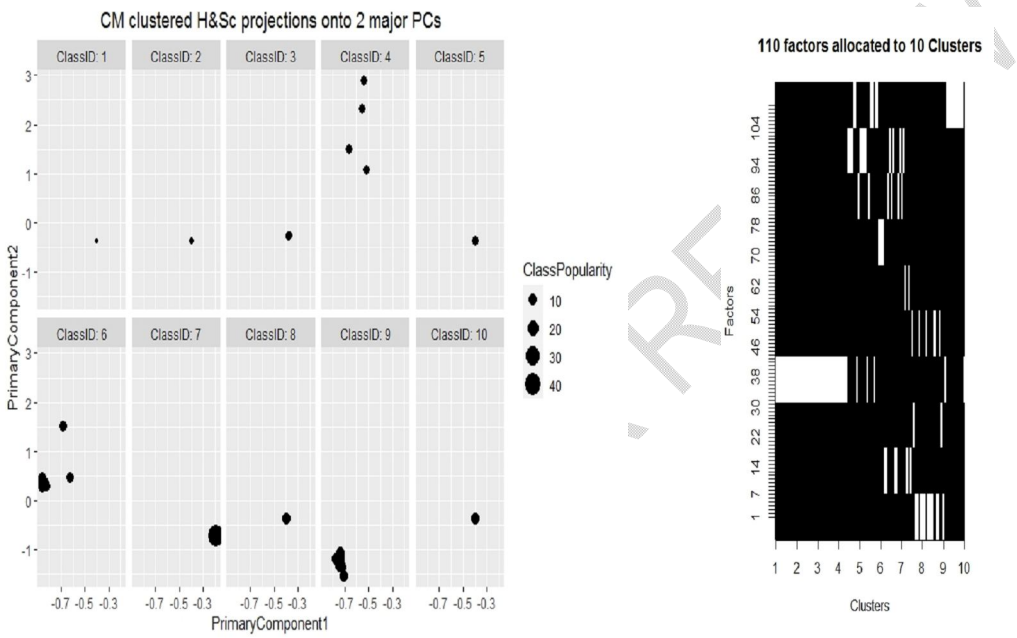


(c)

Fig.3 Plots for indicative linear connections found. The captions below the individual plots show what are the independent variables , (3a) '*Percent of births in Low birth weights* ' and '*Home Care Clients in Home Care Group* ', (3b) '*Mental health problems . gender(female)* ' and '*general practitioners(GP) . value* ', (3c) A multi-linear model with 5 dependent H&Sc factors (each for each plot in (3c)) and target '*Single rooms . care home sector(owned mortgage)* '. The pattern for the X-axes {'N ','/ others' } means that the 1st independent factor is drawn sequentially from the set {'*Delayed discharges(DDs) . 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

4.4 Comparing PCA and LR

PCA is known for minimal data representation and LR is a way to model data using other data as their predictors (Liu(2021)). The two approaches were compared as shown in Table 2. PCA suggests, as in Table 3, that the services pack 'S20' has all the PCs that is also confirmed in Table 3 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.



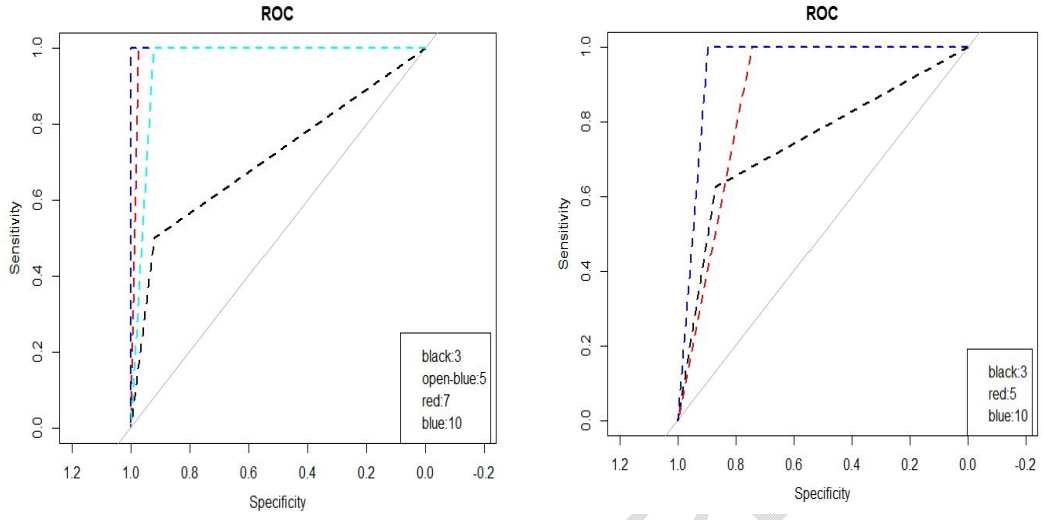
(a) (b)

Fig.4 (4a) The 10 panels represent the 10 clusters from CM while the data are projected onto 2 major PCs which are `PrimaryComponent1` :`*Alcohol-related admissions (stays) or discharges . Alcohol Condition . All Alcoholic Liver Disease (ALD)* ` (S20.A11.L1) and `PrimaryComponent2` :`*Alcohol-related admissions (stays) or discharges . Alcohol Condition . Alcohol Related Brain Damage* ` (S20.A11.L2). The data are represented by circles of a diameter that is proportional to the popularity of the cluster as we see by the overlap of the circles as we move to more crowded panels. The projections are shown for normalized data, (4b) A binary map showing how factors are allocated to clusters (the white rectangles indicate membership and the black non-membership)

4.5 Prediction with different learning/training ratios

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 figure 2. The LR-CM approach can be contrasted to

cluster-wise regression discussed in (Torti, 2019) where the LR coefficients (LR structures) are the meta-data to cluster themselves.



(a) (b)

Fig.5 ROC curves for 4 types of the "RPROP" with (H) layers with (a) {3,5,7,10} nodes for Hsc factor(ID=92) and triplet {S=12,A=13, L=1} which is '*Number GP . Registered Patients . Age . All*', *le.r*=70%. As we add (H) layers the curves move towards the top left corner that is the perfect prediction case, (b) triplet {S=20,A=11, L=4} which is '*Home care.services . Value*'

4.6 ROC anaylsis and NNs

The results from ROC analysis are shown in figure 5 that are obtained using "BPROP" on selected factors in Table 3. The learning algorithms for all 3 algorithms used depended mainly on (H) nodes and less on *le.r.* A deep drop in the RMSE error was observed for (H) layer nodes above 3 and almost at one scale of magnitude. On avarage an RMSE error $Er1 \approx 10^{-1}$ fell to $Er1 \approx 10^{-2}$ as can be seen in the last columns in Table 2, that represent 10 and 15 nodes in (H) layer. Given that the dimension of the data was (39,110), that is, proprtional to the number of (H) nodes this makes sense in terms of well capturing data details.

5. Conclusion

The paper discussed how can we 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 regards 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 and were further explored using LR and ARMA that changed in 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 (Guersel, 2019). 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 specialisation of services offered to alcohol-related patients was confirmed by the high linear confidence attached to such H&Sc factors such as low births weights and services related to alcohol. Depending on the year at hand, though, the '....low birth weight (weight < 2500 gr)' class can also be regressed(linearly related) with mental health patients (Hange U., 2018). 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 that 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.

References

1.Scottish Government(2019). Statistics Service Health and Social Care Data. Available: https ://statistics.gov.scot/data home

2.Vittorio Lippi(2019). Incremental Principal Component Analysis: Exact implementation and continuity corrections". *arXiv: 1901:07922v2*; stat:ML; 13May2019. Available: https ://arxiv.org/pdf/1901.07922.pdf

3.Ian Litchfield(2019). "Can pathways of patients with long-term conditions in UK primary care? A study protocol". *BMJ Open*, 2018. Available: https : //bmjopen.bmj.com/content/8/12/e019947

4.Dimitris Bertsimas(2018), Colin Pawlowski, Ying Daisy Zhuo(2018). "From Predictive Methods to Missing Data Imputation: An Optimisation Approach". *Journal of Machine Learning Research*, 18(2018),1-39. Available: https://jmlr.org/papers/volume18/17-073/17-073.pdf

5.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. Available: http: //dx.doi.org/10.1016/j.compbiomed.2016.06.004

6.de Rooij M.(2018). Transitional modeling of experimental longitudinal data with missing values. *Adv Data AnalClassif,*12,107–130. Available: https://link.springer.com/article/10.1007/s11634-015-0226-6

7.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). Available: https://doi.org/10.1186/s12875-020-01141-w

8.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). Available: https://pubmed.ncbi.nlm.nih.gov/30952873/
9. 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
10. Public Health Scotland(2020). Data and intelligence. A – Z Subject Index. Available: https://www.isdscotland.org/A-to-Z-index/index.asp
11. Capan, 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. Available: https://dx.doi.org/10.4338%2FACI-2015-09-RA-0127
12. 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. Available: doi:10.12788/jhm.3081
13. Guersel, Gueney(2019). Healthcare, uncertainty, and fuzzy logic. Digital Medicine ,2016,2,101-12. Available: https://www:researchgate:net= publication = 310817255Healthcare uncertainty and fuzzy logic
14. 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. Available: https://doi.org/10.1016/j.jval.2014.12.001
15. 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. Available: https ://www.mssanz.org.au/modsim2015/H2/bentovim.pdf
16. 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. Available: doi:10.1016/j.bjps.2014.10.011
17. 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. Available: https://doi.org/10.1371/journal.pone.0180159
18. 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. Available: DOI: 10.1136/bmjopen-2017-018628
19. Harutyunyan, H., Khachatrian, H., Kale, D.C. et al.(2019). Multitask learning and benchmarking with clinical time series data. Sci Data,6,96(2019). Available: https://doi.org/10.1038/s41597-019-0103-9
20. 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.; Nguyen Duc T. (eds),6th International Conference on the Development of Biomedical Engineering in Vietnam (BME6); BME 2017,IFMBE Proceedings,vol 63,Springer Singapore, Available: https://link.springer.com/chapter/10.1007/978-981-10-4361-1 138
21. 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). Available: https://doi.org/10.1038/s41598-020-73740-7
22. 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. Available: DOI:10.3758/s13428-015-0611-2

23. Skiera, Bernd & Reiner, Jochen(2018). Regression Analysis. Homburg, Christian, Klarmann, Martin, Vomberg, Andreas (Editors). Handbook of Market Research. Available: https ://doi:org=10:1007=978-3-319-05542-8_17-1
24. Ciprian Florescu & Christian Igel(2018), RESILIENT BACKPROPAGATION (RPROP) FOR BATCHLEARNING IN TENSORFLOW". Workshop track - ICLR 2018, Available: https://openreview.net/pdf?id=r1R0o7yDz
25. 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). Available: https://doi.org/10.1186/s12962-021-00322-3
26. 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). Available: https : //doi.org/10.1186/s12911/019/0870/4
27. 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(3),e190348. Available: DOI: 10.1001/jamanetworkopen.2019.034
28. 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 . Available: https://arxiv.org/abs/2003.02218
29. Liu C, Zhang X, Nguyen TT, et al(2021). Partial least squares regression and principal component analysis: similarity and dierences between two popular variable reduction approaches. General Psychiatry,2022;35:e100662(2021). Available: doi: 10.1136/gpsych-2021-100662

Table 1. Services and patients cohorts acronyms("S"s) with dates and characteristics per service. Also shown indicative breakdowns of H&Sc packs/data frames into their attributes (the "A"s) and of their attributes into their levels (in parentheses). Where there are no specific attributes as in ("S3") then the 'value ' attribute is used. Many attributes like ages, gender are shared among H&Sc groups.

H&Scs full names(years[3])	A[1]	H&Scs full names(years)	A	Attributes(years)	A
Smoking prevalence in young.people(SALSUS) (1998-2010)	S1	Headcount of general practice workforce (2007-2017)	S15	Weight Category "Epidemiological"(Healthy Weight", "Epidemiological – Obese", "Epidemiological – Overweight", "Epidemiological - Overweight & Obese","Epidemiological – Underweight")	A6
Smoking prevalence and deprivation(SALSUS)(1998-2010)	S2	BMI distribution of primary 1 education children (2001-2019)	S16	Home Care Client Group("Dementia", "HIV, Aids, Alcohol Or Drugs", "Learning Disability","Mental Health Problems","Other Client Groups", "Physical Disabilities")	A7
Smoking behaviour and self rated health(SALSUS)(2008-2019)	S3	Delayed discharges: monthly census (2001-2019)	S17	Care home sector / Type of enure	A8
Primary1 Children Body Mass index epidemiological (2005-2009)	S4	Delayed discharges - Monthly Bed Days Occupied (2016-2020)	S18	Household Type("Adults","All", "Pensioners", "With Children")	A9
Primary 1 Children BMI clinical (2005-2009)	S5	% of children classed healthy weight, overweight, obese,severely obese at Primary 1 review (2002-	S19	Limiting Long term Physical or Mental Health Condition	A10

		2015)			
Places in single rooms in care homes (2002-2011)	S6	Alcohol-related admissions (stays) or discharges (1981-2019)	S20	Birth Weight("Live Singleton Births","Low Weight Births")	A11
Places in care homes (2007-2017)	S7	Drug use among 13 and 15 year olds in Scotland (2002-2015)	S21	Home Care Provider(Value), Headcount of GP(Value)	A12
Occupancy rate in care homes by type of provision, (1996-2018)	S8	Health care clients (2016-2019)	S22	BMI distribution of primary 1 education children	A13
Number of general practices (GPs) with registered patients (2005-2009)	S9	**Attributes(levels)**		2 Or More Clients In Home	A14
Mental wellbeing by tenure, household type, age, sex, disability (2000-2019)	S10	Age(13,15,"All"),Gender(M,F,ALL) A(1,2)	A(1,2)[2]	Home Care Client Group,Delay("Over three days", "2 Weeks +", "4 Weeks +", "6 Weeks +", "All Delays")	A(15,16)
Intensive home care (2014-2017)	S11	Smoking Behaviour("Non-smoker" "Occasional smoker" "Regular smoker")	A3	Delayed Discharges - Monthly Bed Days Occupied	A17
Home care services(2007-2019)	S12	Age Bands("16-34","35-64", "All","16-64", "65 years and over ")	A4	Age bands("16-34" ,"35-64", "All","16-64", "65 years and over")	A23
Living arrangements for home care clients (2007-2017)	S13	Delayed discharges: monthly census	A5	Main Client Group In Care Home("All Adults", "Adults with Learning Disabilities", "Adults with Mental Health Problems", "Adults with Physical Disabilities", "All Adults","Older People Aged 65 and Older", "Other Groups")	A29
Home.care.client.Group (2007-2017)	S14	SIMD quintiles("1 - most deprived", "2", "3", "4"), Weight category ("epidemiological")	A6		

[1]: shorts for names, [2]pairs of attributes, for example (A1,A2) are denoted shortly as A(1,2) to fit the table, [3]years before zero padding

Table 2. Representative results for prediction using (a) LR (TS1 = LR0 + LR1 *TS2 + LR2*TS3 +....), (b)ARMA (TS = AR0 + AR1 * TS(t- 1) + AR2 *TS(t- 2) +...) 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-'), shown are the LR probabilities p1,p2,p3, the LR coefficients(LR0,LR1,LR2) and error metrics RMSE(Er1), MAE(Er2), MRE(Er3), across the 39 years, factors naming follows the attributes levels and h&sc names conventions defined in table 1 and in the figure 1. The NNs are compared for 3 structures (3,10,15 nodes in the hidden layer)

Linear groups of H&Sc factors 1 IDs 2

LR				ARMA			NN(3 layers)	NN(10 layers)	NN(15 layers)
Er1	LR0	P1	Er2	Er1	AR0	Er2	C1(Er1)	C1(Er1)	C1(Er1)
	LR1	P2	Er3		AR1	Er3	C2(Er1)	C2(Er1)	C2(Er1)
	LR2	P3			AR2		C3(Er1)	C3(Er1)	C3(Er1)
(1) S1.A1.L3(Smoking prevalence in young.people(SALSUS) . Age . All), (2) S3.A2.L3(Smoking behaviour and self rated health.(SALSUS) . Gender . All), (3) S1.A5.L2(Smoking prevalence in young.people(SALSUS) . SIMD quintiles . 2)									
0.167	0.016	0.993	0	8.852	-2.70E-003	0	0.372	0.159	0.159
	0.006	0.013	1		-2.70E-003	1	0.406	0.022	0.029
	-2.00E-004	0.04			8.00E-016		0.371	0.02	0.166
(1) S10.A23.L1(Smoking prevalence in young.people(SALSUS) . Age . 13), (2) S10.A29.L1(Mental wellbeing(SSCQ) . Age band . 35-64), (3) S12.A8.L1(Home care services . 2 Or More Clients In Home Value)									
0.945	25264	0.04	0	7609	0.93	0	0.32	0.343	0.0429
	40.6	0.001	1		-7.30E-003	1	0.322	0.08	0.385
	184	0.049			37797		0.322	0.059	0.039
(1) S2.A1.L1(Smoking prevalence and deprivation(SALSUS) . Age . 13), (2) S10.A29.L1(Mental wellbeing(SSCQ) . Age band . 35-64), (3) S12.A6.L4(Home care client living arrangements . 2 Or More Clients In Home)									
0.334	-13.6	0.366	0	95.5	-2.70E-002	1	0.321	0.0216	0.0259
	0.042	0.002	1		-2.70E-002	0	0.323	0.026	0.029
	0.263	0.026			15.6		0.321	0.027	0.032
(1) S2.A5.L1(Smoking prevalence and deprivation(SALSUS) . SIMD quintiles . 1 - most deprived), (2) S10.A2.L1(Mental wellbeing(SSCQ) . Gender . All), (3) S12.A6.L4(Home care client living arrangements . Alone), (4) S3.A2.L1(Smoking behaviour and self rated health (SALSUS) . Gender . All), (5) S20.A5.L2(Smoking prevalence in young.people(SALSUS) . SIMD quintiles . 2),(6) S10.A9.L1(Mental wellbeing(SSCQ) . Household type . Pensioners)									
0.483	2.361	0.024	0.025	7.605	-0.3	230	0.321	0.141	0.0259
	0.0173	0.096	1		0.99	4	0.316	0.159	0.158
	4.00E-003	0.213			0.195		0.316	0.028	0.148
(1) S20.A11.L4(Alcohol-related admissions (stays) or discharges . Alcohol-related admissions (stays) . Stays), (2) S3.A2.L3(Smoking behaviour and self rated health(SALSUS) . Gender. Male), (3) S20.A11.L2(Alcohol-related admissions (stays) or discharges . Alcohol-related admissions (stays) . New Patients)									
0.167	0.016	0.993	0	8.852	-2.70E-003	0	0.406	0.224	0.255
	0.006	0.013	1		-2.70E-003	1	0.397	0.34	0.313
	-2.00E-004	0.04			8.00E-016		0.397	0.361	0.358
(1) S3.A4.L3(Smoking behaviour and self rated health(SALSUS) . Self assessed general health . Good), (2) S20.A2.L1(Alcohol-related admissions (stays) or discharges . Gender . Male), (3) S4. A29.L3(Primary1 Children Body Mass index epidemiological . Age Bands . 75 years and over)									
0.821	0.024	0.026	1	0.98	-0.08	0	0.396	0.158	0.159
	0.024	0.035	119.5		70	1	0.398	0.158	0.159
	1.00E-003	0			63		0.398	0.157	0.139
(1) S12.A13.L1(Number GP registered patients. Age. 65 and over), (2) S12.A1.L1(Number.GP registered patients . Age .16.64), (3) S12.A1.L1(Number.GP registered patients . Age . All)									
0.049	1.654		0.608	2.10E-008	0.962	10309	0.35	0.214	0.183
	-6.53	0.025	1		-0.061	0	0.351	0.217	0.243
	-0.002	0.046			-0.041		0.356	0.217	0.243
(1) (S10.A23.L1)Mental wellbeing SSCQ . birth weight. Low. weight births, (2) (S14.A2.L3)Home intensive . Home care . Gender. Male, (3) S17. A5. L3(Delayed discharges: monthly census . Smoking behaviour . Regular smoker), (4) (S7,A2,L3)Primary 1 children . BMI epidemiological . Gender . Female)									
0.349	0.08	3.00E-015	1.00E-023	929.98	0.705	0	0.391	0.039	0.043
	6.89	0.999	1		-9.60E-004	271	0.403	0.097	0.104
	-0.005	0.968			-0.261		0.4	0.251	0.395
(1) S5 . A1. L3(Smoking prevalence and deprivation(SALSUS) . Gender . Male), (2) S3. A4 . L5(Smoking prevalence in young people(SALSUS) . SIMD quintiles 5...least.deprived) , (3) S3 . A4. L1(Smoking prevalence in young people(SALSUS) . SIMD quintiles 5...least.deprived)									
0.998	24.2	0.615	0.45	2.00E-014	1	-0.03	0.391	0.0184	0.027
	0.406	0.675	0.83		-0.03	1	0.403	0.025	0.158
	-0.139	1.00E-023			-0.03		0.4	0.025	0.16

(*1) H&Sc factors are shown as triplets x:y:z (c.f. Section "C"), (*2) attributes and levels as per Table II, 1st and 2nd columns, 3 naming convention not shown(but used), 4 are the levels of attributes (some are listed in Table I)

Table 3. Best Hsc factors using PCA. The factors are represented by triplets {X.Y.Z} as defined in Table 1. The dominant services are "S2"("*Smoking prevalence and deprivation(SALSUS)*") and "S20" ("*Alcohol-related admissions (stays) or discharges*")

Factor name	PC(%)	Factor name	PC(%)	Factor name	PC(%)
Smoking prevalence in young people(SALSUS) . Age. 13(S1.A1.L1)	69	Smoking behaviour and self rated health(SALSUS) .Gender. Female(S3.A2.L2)	14	Smoking behaviour and self rated health(SALSUS) . Self assessed general health . Fair (S3.A4.L4)	6.46
Smoking prevalence in young people(SALSUS) . Age. 13(S1.A1.L2)	5.07	Smoking behaviour and self rated health(SALSUS) .Gender. Male(S3.A2.L3)	3.01	Smoking behaviour and self rated health(SALSUS) . Smoking behaviour . Non Smoker (S3.A3.L1)	0.7
Smoking prevalence in young people(SALSUS). Age.All(S1.A1.L3)	0.59	Smoking behaviour and self rated health(SALSUS) . Self assessed general health . Very good (S3.A4.L5)	0,32	Smoking behaviour and self rated health(SALSUS) . Smoking behaviour .(S3.A3.L2)	0.29
Smoking behaviour and self rated health(SALSUS) .Gender. All(S3.A2.L1)	0.11	Smoking behaviour and self rated health(SALSUS) . Self assessed general health . Bad (S3.A4.L1)			

Consent (where ever applicable)

The paper uses open public PHS data and a reference to NHSS or to PHS is in the references.