Strategies to Minimize Data Sample Size for Regression-Based Pump/Motor Models

Jack L. Johnson[1]*, Jose Garcia-Bravo,[2] Pawan Panwar[3] and Paul Michael[4]

[1]IDAS Electrohydraulics, Waukesha, WI., U.S.A. *Corresponding author
[2] School of Engineering Technology, Purdue University, West Lafayette, IN., U.S.A.
[3] Department of Mechanical Engineering, University of California – Merced, Merced, CA., U.S.A.
[4] Fluid Power Institute, Milwaukee School of Engineering, Milwaukee, WI., U.S.A.
E-mail: jack@idaseng.com, jmgarcia@purdue.edu, ppanwar@ucmerced.edu, michael@msoe.edu

Abstract

This work presents an analysis for tracking the evolution of regression coefficients and the Root-Mean-Square of their residuals on a test dataset for a hydraulic pump. The method starts by iteratively regressing data points that are undergoing sequencing by adding one new data sample at a time, then regressing with each iteration. This process was named Progressively Sequenced Regression Analysis, shortened to "PSR analysis" in this paper. The motivating and guiding postulate of PSR analysis is based on the belief that a plateau of the regression coefficients and statistical figures of merit had to exist if sampling theory is accepted to be real. It was anticipated at the outset that both the regression coefficients and the Residual RMS would converge on respective plateau values; however, it was discovered that the coefficients were very volatile, with some, more volatile than others. Tracking the Residual RMS was found to produce the more reliable measure of information saturation because the convergence is more obvious, provided that the sample sequencing was done with the experience learned from performing PSR analysis. This document is focused on explaining how orthogonally sequenced data can be mined for the limits or hyperspace vertexes of the sampled data, and the source data optimally sequenced (rearranged) to produce results that are as efficacious as Latin Hypercube (LHC) sampling for achieving information saturation at a predictable number of samples. PSR analysis has led to an objective method for verifying that the proper arrangement, i.e., optimized sequencing, of the source data set can predict the condition of information saturation and minimum useful sample size. It ends with a postulate of how this can be achieved using a combination of LHC sampling and vertex pre-test planning, or vertex mining of legacy data. The content of this paper has concentrated solely on the output flow model of hydraulic, positive displacement pumps.

Keywords: RMS of Residuals, Progressively Sequenced Regression Analysis, Latin Hypercube sampling, minimum sample size, vertex mining, vertex sequencing, hyperspace vertexes, convergence plateau, pump flow model

1 Introduction

Hydraulic pumps and motors are widely used in industrial and mobile applications requiring relatively high torque and shaft speed. In many systems, these components are preferred for their high power density, flexibility and simple operation. Unfortunately, despite being ubiquitous in countless applications, the average hydraulic system using pumps and motors presents efficiencies as low as 26% [1]. A solution to improve the efficiency of such machines is to develop accurate and reliable models that predict the behavior of pumps, valves, conductors and actuators. These models are used in industry to develop and engineer highly productive and efficient hydraulic systems through the selection of meticulously analyzed components. However, these systems can only be effective if the models are reasonably accurate.

The means for modeling hydraulic pumps and motors vary in the reported literature. Concerning hydraulic pumps and motors, the most notable means for describing the characteristic performance of a pump correspond to the

flow model and the torque model. For an ideal pump or motor, the displaced fluid is proportional to the velocity of the shaft speed, eq.1 shows this relationship.

$$Q_T = V \cdot N \tag{1}$$

Where, Q_T is the theoretical (ideal) flow, V represents the volumetric displacement of the pump or motor, and N is the shaft speed. Various methods are found in the literature for obtaining the volumetric displacement of a pump or motor [2]-[5]. And all of these seek to determine the derived displacement, a proportional constant for estimating the flow output of a pump at a given shaft speed. More sophisticated methods like the one presented by Eggers, et al. [6] were used to describe a mathematical interpolation procedure called POLYMOD, which was used for developing a torque and flow model based on a polynomial fit of experimental test data. Similarly, Conrad et al. [7] developed a loss model where the flow losses could be obtained from fitting experimental data to a line. Whether the aim is to obtain a complete flow model or simply for determining the volumetric displacement of the pump or motor, a set of experimental data points is required. However, determining how accurate the chosen method is, remains a question of not only the principles on which the method itself is based, but also the quality and volume of the data, that is, the number of experimental data points needed to obtain a useful model. The aim of this work is to present a methodology for determining an optimal sample size for the creation of models based on experimental data, which in turn, will improve testing performance. The motivation for this project was to reduce the amount of time on the test stand while at the same time increasing the amount of production hardware that would be subjected to complete mathematical modeling. Original Equipment Manufacturers (OEM) such as manufacturers of agriculture, construction, forestry equipment, etc., and researchers increasingly expect component manufacturers to supply accurate performance information and models of their components for their own system simulations and performance predictions. The single most important criterion for successful and complete creation of minimal and efficacious test data is to collect the samples in the proper order. The following sections explain how to achieve proper sampling order, and a sequenced regression method presented in this work, herein described as PSR analysis needed to estimate the number of test samples.

1.1 Overview of the PSR Procedure

Progressively sequenced regression analysis was first introduced by Johnson [8] as a method for verifying that a given number of test samples would be sufficient to reach information saturation for the creation of a mathematical pump or motor model. That study explained how Latin Hypercube (LHC) [9] experiment design was crucial to discovering the importance of data order. He also made the case for using PSR analysis as a research method that could lead to minimization of sample size for modeling purposes. Earlier Progress Reports by Johnson to the International Organization for Standardization [10]-[12] showed how the success or failure of PSR analysis is a matter of the order in which the samples are regressed. When sample order is arranged properly (skillfully), information saturation is revealed reliably, but more importantly, the number of samples needed to reach that saturation is controllable through the order used to collect the data in the laboratory, or alternately, the post-lab reordering of the data by optimal analysis, such as with data mining and vertexing.

The PSR process begins with a thoroughly ordered and nominally large source dataset called in this document the genesis file. The regression starts with only the first few samples taken from the genesis file and, placing them into a matrix to be regressed called the PSR matrix. The PSR matrix is subjected to evaluation using an ordinary, linear, multiple regression program to fit test data to a pump or motor model. eq. 2, below presents a simple flow model to be used for the linear regression, this model contains only three regressor terms, and was used in the study at hand.

$$Q_T = A_0 + A_1 N + A_2 \frac{p}{\nu} + A_3 N p \tag{2}$$

Where Q_T represents the theoretical flow estimation for this proposed mathematical model, A_n correspond to the regression coefficients for this model, N is the pump shaft speed, p is the pump differential pressure and ν is the kinematic fluid viscosity This model is adapted from the flow model first presented by Toet [5]. Other pump or motor flow equations may be used for the regression, allowing for performing studies on the effect of other operating parameters or combinations of them. Other, more complex functions were used, such as the one for fig. 2 and fig. 3. and are shown for reference.

After the regression algorithm was evaluated with the first sample points from the dataset, the regression coefficients and chosen figures-of-merit (FoMs) were stored in an output file for later viewing and processing. Examples of FoMs include; sample mean, sample standard deviation, RMS error, etc. Each iteration of the PSR

analysis added one more observation to the output data file, which contained all the regression coefficients and chosen FoMs. In this way, the evolving coefficients for the math model and FoMs could be tracked as the number of samples grew by one sample per iteration. The term "information saturation" is defined as the point at which a sufficient number of samples was reached. It was observed that when information saturation was attained no substantial changes in the RMS Residual Error followed after adding samples and iterations.

Skillful or optimal selection of the data sequencing means that the order of the first samples in the original experimental dataset or genesis file, have been selected by means that are described in this paper in the subsection below. The aim is to demonstrate that the data set can be reduced to bring about information saturation with the fewest number of data samples and furthermore, the number of samples needed to obtain information saturation is predictable as a result of implementing those optimized methods. Some details of these procedures are discussed in later sections of this paper.

1.2 The Progressively Sequenced Regression Algorithm

Figure 1 shows the basic PSR analysis flow chart that outlines the overall procedure of Progressively Sequenced Regression Analysis. The authors' algorithm used multivariate linear regression in an iterative process, where a new data sample from the genesis file is added with each new iteration until all samples in the genesis file were processed in the analysis.

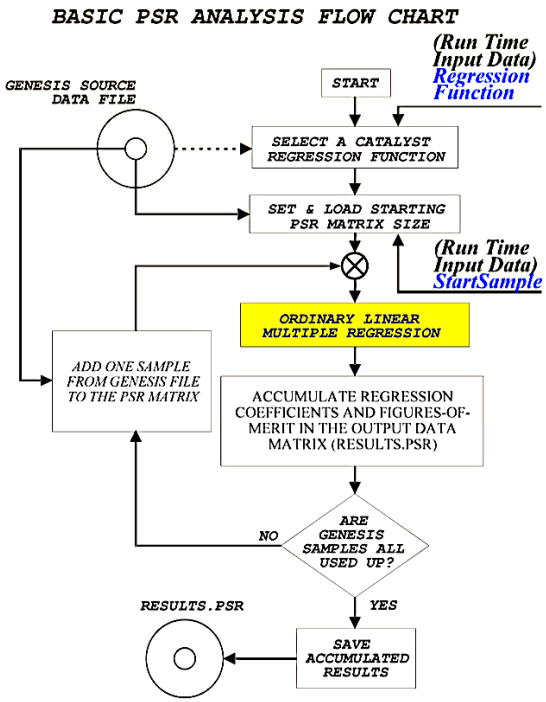

Figure 1: Basic flow chart of the PSR algorithm

Specifically, once the PSR algorithm is fed the input data, two other essential pieces of Run Time Input Data are required by the PSR analysis software. The algorithm begins with the Regression Function corresponding to any form as exemplified in eq. 2. In the first step the desired regressor columns are picked from the genesis file. In the case study presented here, these regression coefficients correspond to the following physical variables which reside in generated columns in the genesis file (data matrix):

a. Shaft Speed (rad/s) or (rpm), N, for the A_1 term

b. Pressure/viscosity ratio (Pa.s/m^2) or (bar/cSt), p/ν, for the A_2 term

c. Speed · pressure product, (rad/s·Pa) or (rpm·bar), $N{\cdot}p$, for the A_3 term

1.3 Review of previous findings of PSR analysis

Data crowding is a term used to stress the detrimental effects caused by consecutive samples that contribute little change from one sample to the next. They will most likely occur in the opening regions of PSR analysis, that is, when the regressed sample count is the smallest (smallest PSR Matrix). When using the orthogonal sequencing as obtained from ISO 4409 data, the effects of data crowding create problems in PSR analysis. Changes in measured speed, for example, at one sample to the measured speed at the next sample are small. The changes in the speed variable are so small as to create a near zero value of the regression determinant and will result in coefficient values that are very large and changing rapidly from data sample to data sample (coefficient volatility). Such effects are apparent in the estimation of A_0 coefficient for the regression function as shown in fig. 2.

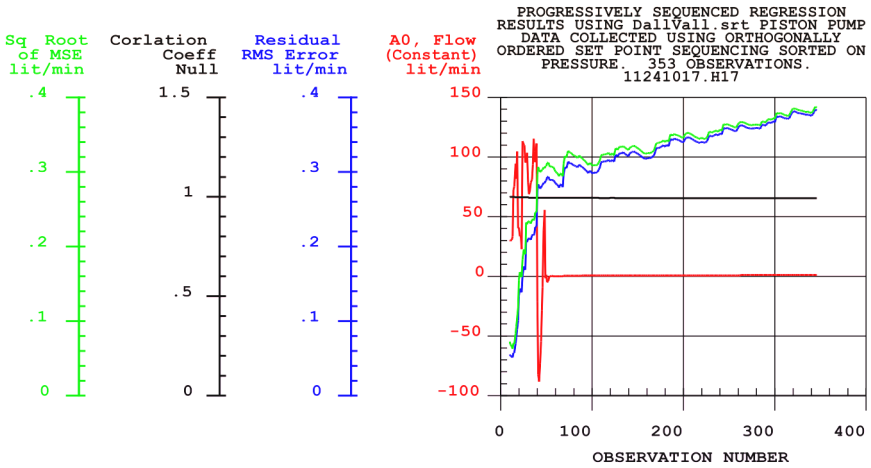

Figure 2: Example of failure to reach information saturation and coefficient volatility for an axial piston pump.

Figure 2 shows that the A_0 coefficient has gone "flat line" at just over 50 observations, suggesting that information saturation has been reached. This is a false positive because it is the extreme swings in A_0 in this region that expand graph scaling such that information saturation appears to have occurred. The existence of the flat region is a result of graphical scaling and not true arrival at the convergence zone. The failure to reach convergence becomes apparent only after the most volatile region is suppressed as is shown in fig. 3. Suppression was done simply, by setting PSR's internal StartSample variable to 52 samples at the opening of this example PSR analysis.

Setting the SampleStart internal variable to a larger value does not ignore or discard valid data, it merely suppresses the volatile output data caused by those samples and thus avoids the graph scaling problem caused by volatility in A_0. This becomes apparent in fig. 3 in which the volatility of the first 52 data samples has been suppressed. That is, the very first regression in the PSR analysis (fig. 3) used a PSR matrix with 52 data samples. They are valid data points taken in accordance with an ISO standard procedure, however, putting them first in the inputting sequence created the volatility due to data crowding. It takes 52 regressed samples to "flush out" the influences of the data crowding. After that, the A_0 coefficient is "better behaved" and the output is more meaningful and can be correctly interpreted. Fig. 2 and Fig. 3 show an example of a failed PSR analysis.

It was further observed that data order or data sequencing is paramount and holds the solution to successful PSR analysis. The genesis file used to generate the PSR analysis output of fig. 2 and Fig. 3 was subjected to a conventional ordering on the independent variable, differential output pressure. Early in the research, while trying to determine the processes that might control the ability to reach information saturation, sorting was done on pressure to assess the effects. The sorting based on pressure was done in ascending order, resulting in all of the lowest pressures occupying the first data samples right at the beginning of the genesis file. Along with orthogonal sequencing and its accompanying data crowding per the recommendations provided in ISO 4409, there was no worse way to prearrange the data. Sorting using the pressure was a clear way to invite data crowding problems in a worst case way, fig. 2 demonstrates this fact, and fig. 3 reinforces it. Covered in later sections of this paper are the discoveries that Latin Hypercube sequencing and orthogonal data vertexing can result in a minimization/reduction of number of samples needed for useful models because they give to the regression process the extremities of the tested hyperspace in the earliest samples, assuring attainment of information saturation.

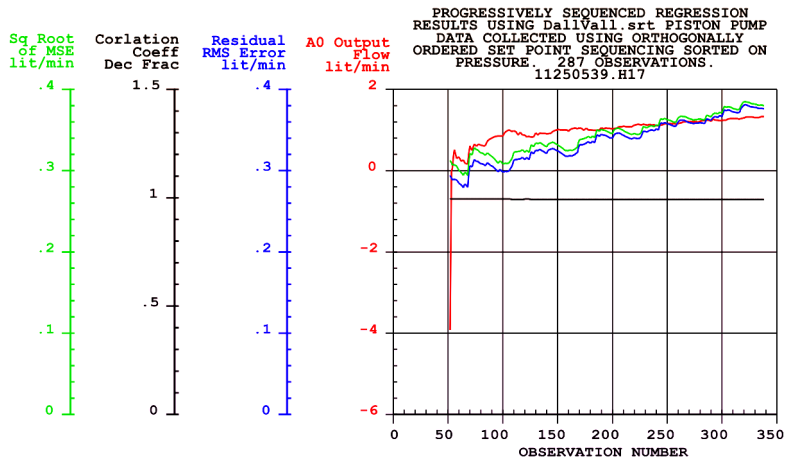

Figure 3: Detail of the A_0 coefficient for an axial piston pump without first 52 samples.

2 Methodology

The plots shown in fig. 3 contain the same curves as in fig. 2, but with the A_0 volatile data removed from the display. The reader is urged to observe the similarities in the curves of fig. 2 and fig. 3 by first seeing the large differences in the scaling of the A_0 axes in the two figures. In fig. 2 the A_0 spans from -100 to +150 lt./min (-0.00167 m^3/s to 0.0025 m^3/s) while the range in fig. 3 is from -6 to +2 lt./min (-0.0001 to 0.000033 m^3/s). In fig. 3 the most volatile range of A_0 data has been removed from the display for the purpose of exposing A_0's real trends with the volatility removed. The similarities in the two graphs lie in the other three variables beyond 52 observations. They have the same scaling and the same shapes in the two respective figures. It is clear in both figures that information saturation was never reached because the Residual RMS Error the Square Root of MSE both rise steadily as the PSR analysis sample count increases toward the end value of 353 samples. There is no convergence plateau. This, too, is a direct consequence of the sorting of the genesis file using pressure as the variable. Later sections of this paper show that the data can be re-sequenced so that information saturation does occur, and the number of samples needed to reach it can be controlled. Likewise, PSR analysis reveals many anomalies of regression analysis. For example, there is likely to be more volatility in the individual coefficients A_0 shown in fig. 2, than there is in the Residual RMS Error leading the researchers to adopt it as the FoM of choice.

PSR analysis is a way to exhibit the volatility graphically. Users of regression might not otherwise be aware of this volatility. Unfortunately, regression cannot extract coefficients from data with the same certainty of, say, Fourier analysis can extract harmonic components from a cyclic waveform of arbitrary shape. A given regression coefficient value depends upon the makeup of other companion regressors in the same catalyst regression function. In other words, the coefficient volatility is affected by the form of the catalyst regression function, but information saturation is affected by the order in which samples are presented to regression.

For many modelers the conventional FoM for the quality of a model is the R^2 value. In output flow modeling of a positive displacement pump, the nominal uncertainty in the independent and dependent variables is about ±0.5%, subjectively speaking. This means that the total uncertainty in a model is probably in the range of ±1% or ±2%. This means further, that about 98% of the pump performance is predictable by that amount. The random variations are only one or two percent. But it also means that the discrimination of one model to the next with the R^2 value is in the 6^{th} or 7^{th} decimal place, making R^2 a poor discriminator of model quality. Residual RMS Error is a far better discriminator, a reality that led the authors to adopt it as the FoM of choice.

Additionally, and just as importantly as selecting the correct FoM is the fact that sequencing must be such that the extremities of the tested hyperspace (vertexes and/or Latin Hypercube experiment design) must be present in the earliest samples in order to achieve information saturation. Any future data points that are beyond the hyperspace limits of past data points is probably going to create a rise in the Residual RMS Error and lead to a new, higher plateau. The remainder of this paper reports on the means by which the outer limits of the testing hyperspace can be controlled and information saturation can be reached.

2.1 Preparation of the dataset using an LHC experiment design

LHC sampling was used by Panwar and Michael [13] to create a designed sampling experiment using the sampling strategy followed by McKay et al. [9], their goal was to build a set of test set points to guide and sequence data collection in the laboratory. Johnson [8] used this data to carry out a PSR analysis. The results of the analysis revealed unambiguous information saturation that coincided almost perfectly with the number of samples chosen for the Latin hypercube data set (25, 50 and 100 samples), which was set to be first in each of the several genesis files and thus became the PSR learning zone as suggested in fig. 4. The learning zone is a term used in this work to explain and separate that early evolutionary region in PSR analysis where the Residual RMS Error has not yet settled on a convergence value. It is introduced graphically in the idealized graph of fig. 4. Not all data behaves this way, but is similar when sequencing is enhanced to achieve information saturation. The quest of the researchers was to find data sequencing that produced similar outcomes and unambiguous information saturation in the evolution of the FoM. Johnson [8] observed that the results produced an optimized minimum sampling requirement when using the Latin Hypercube test plan because the information saturation was obtained with a lower sample count. In this document finding this minimum requirement is referred to as optimality.

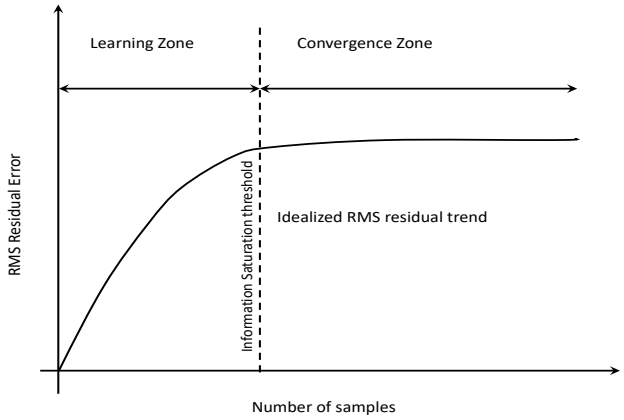

Figure 4: Ideal trend of the RMS value for a regression model.

2.2 Input Data for the LHC Algorithm

The data for the PSR analysis for this work was obtained from testing a hydraulic pump and following the standardized test procedure described in ISO 4409 [14]. The test set point sequence consisted of the aforementioned Latin Hypercube designed experiment of 25, 50 and 100 samples respectively, to serve as learning zones in several test sequences. These learning zones were followed by generally larger LHC sequences to serve as convergence zones with assurances that the hyperspace limits were the same in all LHC sequences regardless of being positioned for a learning zone or for a convergence zone, an absolutely necessary requirement as the experiment was designed using Matlab®. The subject test specimen was a variable, positive displacement hydraulic pump producing an approximate flow, Q, of 100 lt/min (0.00167 m^3/s) at the maximum test speed, N, of 2200 rev/min (230 rad/s). The test set point sequence consisted of the specified number of data samples to be used as test set points in the laboratory and were constrained by the specified ranges given for the three independent test variables shown in eq.1, output pressure p, shaft input speed N, fluid viscosity ν. The LHC algorithm function from Matlab® was used for creating pre-laboratory test set points for the independent variables to be measured. Whether a test set point sequence is a learning zone or a convergence zone is a matter of placement in the genesis file plus the successful reaching of a convergence plateau.

After implementing the laboratory test plan which produced the measured data, and followed by calculating candidate independent regressors, a viable genesis data file was built, from which PSR analysis could then be performed. It was the integration of LHC sampling verified with PSR [8] that revealed the efficacy of such sampling.

Optimality was obtained in all cases because the number of samples specified in the creation of the LHC learning zone predicted almost perfectly the number of samples needed to reach information saturation. Current research now points to inclusion of hyperspace vertexes along with LHC sampling to create a more effective and optimally improved learning zone. This will in turn yield a better representation of the population of test data.

2.3 Definition of learning and convergence zones using LHC sampling

The LHC algorithm had to be invoked at least two times to create a single designed experiment: once for the "learning zone" and a second time for the expected "convergence plateau" to appear during PSR analysis iterations [8]. The learning zone samples form the first part of the test plan and the convergence plateau samples follow, and they must be in that order. If, after PSR has analyzed those learning zone samples, the RMS Residual Errors remain nominally constant, then that is a necessary and sufficient condition to confirm that information saturation has occurred (fig. 4). No instances of actual constant RMS residual error in the convergence plateau were reported. Instead, and as expected, the error varied slightly in the convergence plateau, however, Johnson [8] did not recommend objective estimates of acceptable variation limits. It is postulated now, in this paper, that a reasonable limit in the plateau zone be set equal to, or less than, the allowable measurement error that applies to the dependent variable being modeled, or a specified fraction thereof. It would be a starting point, and an assistance to the objective determination of minimum sample size for accurate models.

2.4 Definition of the start of the convergence zone

Figure 5 contains the results of the three different PSR analyses with test data from a single axial piston pump. The designed experiment explored the consequences of using LHC sampling with the three different learning zone sample counts in order to see the effects on the number of samples used, or PSR iterations, needed to reach information saturation. The terms "25's Curve", "50's Curve" and "100's Curve" in fig. 5 refer to the three genesis files with 25, 50 and 100 LHC sample sets as learning zones, with each followed by larger convergences zone sample sets.

However, this was accomplished by manually shifting data samples so the total number of samples was always 300 and all 300 were the exact same total set, but with the leading samples being comprised of the 25, 50 and 100 LHC samples respectively.

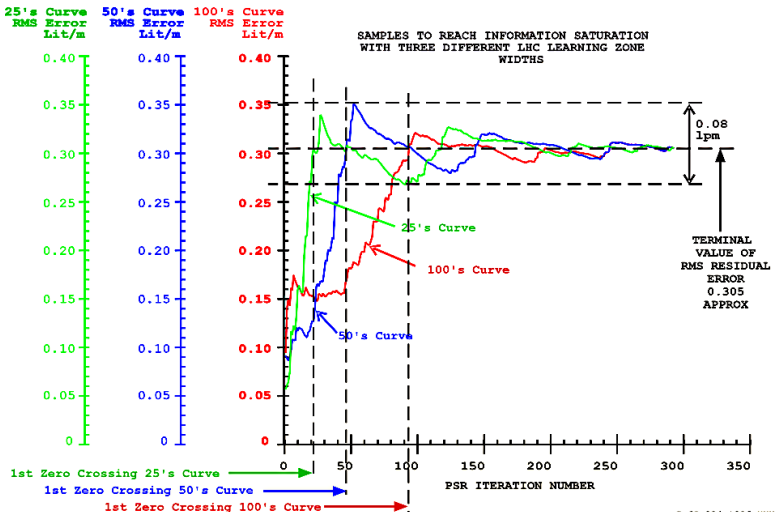

Figure 5: RMS error of predicted flow for various learning zone sample sizes.

With the semi-random nature of LHC sampling, this was the only fair way to compare the efficacy of the three learning zones. This explains why all three graphs converge on the same final value of Residual RMS Error. The criterion for having reached information saturation is the crossing of the final value of the RMS Residual error that is reached after a much longer convergence zone. fig. 5 identifies this final value as approximately 0.305 lt/min (5.083×10^{-6} m^3/s) for the chosen catalyst regression function presented in eq.2. An exact mathematical function to model a pump is not known, however, it is known that experimentally validated models can have 6 variable terms plus the constant [13]. This means there were seven unknown coefficients, thus the regression needed a minimum of eight samples as the absolute minimum. PSR analysis started with the eighth sample. This starting sample count, StartSample, corresponds to a PSR iteration count of one. fig. 5 is constructed with PSR iteration count as the horizontal axis so the first entry on the graph occurs at the value of 1. The iteration count in fig. 5 is what statisticians refer to as "degrees of freedom" which takes into account the number of unknowns in the catalyst regression function, however, in general, the iteration count can be independently controlled when PSR analysis starts in order to achieve some special graphical display effects and as demonstrated in comparing fig. 2 and fig. 3.

The designed experiment was used to generate the convergence zone samples as well as the learning zone samples. For all three analyses in Fig. 5, there was a total of 300 samples, therefore, the graph ends at 292 iterations. The three PSR analyses were carried out with a different number of samples each. The "25's Curve" meaning that its first 25 samples were taken from a complete LHC designed experiment set up to generate 25 designed samples. It led to a first crossing point at approximately 21 iterations corresponding to 28 data points. The "50's Curve" meaning that its first 50 samples were taken from a complete LHC designed experiment set up to generate 50 designed samples. It led to a first crossing point at approximately 46 iterations corresponding to 53 samples. Lastly, the "100's Curve" meaning that its first 100 samples were taken from a complete LHC designed experiment set up to generate 100 designed samples. It led to a first crossing point at approximately 92 iterations corresponding to 100 samples.

Additional data has been identified in fig. 4, notably the final value of the RMS Residual error, which is estimated to be about 0.305 lt/min (5.083×10^{-6} m^3/s). It is evident that the residual error does not reach a constant value, but instead fluctuates. The ultimate limits after first crossing, which take into account not just the peak-to-peak value of one curve, but the composite peak-to-peak of all three curves, is indicated by the span of about 0.08 lt/min (1.33×10^{-6} m^3/s) on the graph of fig. 5. The interpretation of the convergence zone can be estimated as 0.305 ±0.04 lt/min ($5.083 \times 10^{-6} \pm 6.667 \times 10^{-7}$ m^3/s). This is a trivial amount of variation in light of the maximum measured output flow which is just under 100 lt/min (0.00167 m^3/s) for this pump as tested. The authors believe that the minimum number of samples for this pump is 25 samples, that is the learning zone width, provided those 25 samples are collected using a 25 sample LHC-designed test set point sequence. There is every reason to believe, that a fewer number of samples will produce a useful model, provided the LHC, or equivalent, algorithm is used to establish the test's set point sequence. It will be shown shortly that when the learning zone is built using the hyperspace vertexes of an orthogonal test point sequencing, as is dictated by ISO 4409, that the ability to predict and reach information saturation is just as efficacious as is the LHC algorithm. For purposes of this paper, hyperspace vertexes are the upper and lower extremities of the independent variable ranges, that is, no test points exist outside the tested hyperspace, of the independent variables in all their combinations.

It can be concluded that with the LHC design of experiment method, the minimum sample size is an independent variable because modeling of hydraulic pumps and motors is concerned with known physics used for the mathematical model or catalyst function as shown in eq. 2. This model carries an "ideal flow term", $A_1 \cdot N$, and an internal laminar leakage term, $A_2 \cdot p/\nu$, both of which are well known hydromechanical phenomena occurring in hydraulic pumps and motors, plus the so-called Couette/compressibility effect, i.e., the last term in eq 2. fig. 5 shows that the LHC sampling algorithm can produce a designed experiment that will always produce information saturation within the number of samples requested of the LHC algorithm. That is, the points of the first crossings coincide almost perfectly with the numbers of samples used for each respective LHC learning zone width, i.e., 25, 50 and 100.

2.5 LHC Sampling for prediction of information saturation

The authors sought to explain why when using an LHC experiment design, the point of first crossing of the final RMS residual error always coincided regardless of the number of samples selected for reaching the learning zone. It was hypothesized that because LHC is a randomization method for selecting the data, a better sampling of the pump performance can be obtained for the regression used in PSR. Therefore, the following statements can be made:

- Latin Hypercube sampling strategically spreads sample points evenly throughout the chosen data hyperspace.
- The stratification of the data used in LHC layering of the sample points creates assurances that all regions of the hyperspace are fairly sampled,
- Randomization distributes the sample test points throughout the defined hyperspace.

However, the authors recognized that conventional LHC sampling always falls short of the tested hyperspace limits, therefore the hyperspace vertexes became new possible candidates for populating the learning zones. This paper validates the vertexing strategy. Given these statements, LHC is found to be an efficient method for sampling of the entire universe of points within the hyperspace of the pump's performance. PSR analysis verifies adequacy of the sample size by reaching information saturation within the sample count used in the learning zone. In PSR analysis the adequacy of the small sample is referred to as efficacy of sampling. To the best of the author's knowledge, statisticians have not given a name to that phenomenon which allows a small number of samples to serve as a model for a much greater population of samples. PSR analysis using LHC sampling has shown empirically that even sample sizes as small as 25 provide and predict that information saturation will be reached.

LHC sampling has been producing its sufficiency in sampling since McKay first introduced it in 1979 [9], however, the sufficiency was not known until PSR analysis was used to verify it. The virtues of LHC design of experiments have been lauded by other authors [15]-[17] , however, the laudits are not universal. For example, Vose [18] argues against the value of LHC sampling but, does so in the original intent and context of McKay [9]. The original intent was aimed at efforts to minimize the number of iterations needed to draw conclusions from simulations with scores or hundreds of parametric combinations. PSR analysis with LHC sampling in this current context is used to minimize data samples needed to create useful physical models drawn from the data. PSR analysis revealed, and verified, the original postulate by Johnson [8], that there must come a point where additional sample counts contribute little or nothing to the quality of the model.

2.6 The role of rearranging the genesis file

The authors investigated the beneficial effects of randomizing an orthogonally collected genesis file after having failed to reveal information saturation without ambiguity. Randomizing was carried out using a reordering algorithm which in turn made use of the semi-random number generator in their compiler. The results created an unambiguous RMS error plateau as seen on fig. 6, but it lacked the predictability of LHC sequencing. That is, LHC sequencing provided the sample count to reach information saturation, by inspection, while at best, some statistical theories likely have to be applied to make sense of the saturation sample count with random reordering of the data. Nevertheless, this data randomized sorting appears to be reliable in demonstrating information saturation, provided there are a sufficient number of data samples. But the uncertainty in predicting sample count for saturation precluded investing any further time into a search for a predictor. No further investigations were made using rearranging beyond the five corroborating analyses for a fixed displacement gear pump shown in fig. 6 and a variable displacement piston pump fig.5. For the gear pump the ideal shape suggested in fig. 4 emerged in fig. 6 and it leaves no doubt that information saturation occurs at about 30 or 40 samples. In order to use the model associated with the first observations below 30, the sorting of independent variables and the PSR analysis on unaltered orthogonally sequenced data will not work. Randomization of the data will provide models from fewer samples, however, there is no reason to believe that fig. 6 depicts a condition of optimality.

2.7 PSR and optimal sampling efficiency Discussion

In an ideal scenario, measurements would be made with zero error. Of course, a zero-error experiment is impossible, but a discussion of such an ideal scenario helps to point out the challenges in the real world scenario. The requirements for modeling an ideal, perfect, zero-error data set are listed below:

- The smallest possible PSR matrix must have only one more sample than there are unknown regression coefficients in order to satisfy this well-known requirement of regression, and,
- There must be some variation in each and all of the independent variables within the selected M+1 samples, where M is the number of unknown coefficients in the known regression function.

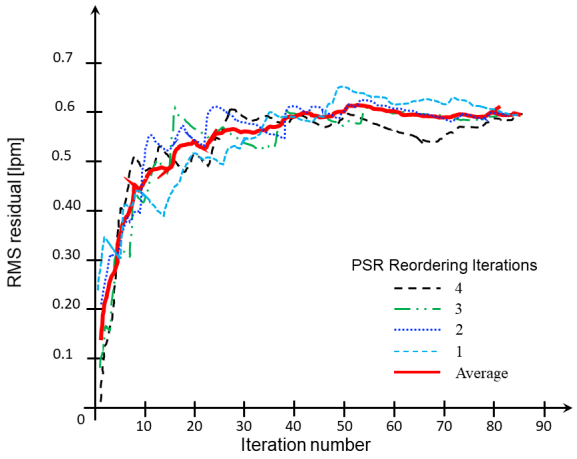

Figure 6: Effect of iterative data rearranging on RMS residuals for a gear pump

In this ideal case, assuming a perfect pump model, M+1 samples are all that are needed, provided both principles are met, then the perfect regression function and the perfectly measured data will absolutely guarantee a perfect set of coefficients. All error-based FoMs i.e., the RMS error, will be exactly 0.0. Perhaps more importantly, if there are more than M+1 samples collected by the laboratory, it makes no difference which M+1 samples are selected. Any M+1 samples, or more, within the entire assortment of collected samples will absolutely produce the exact same coefficients and will absolutely produce zero error between model and measured data. More samples are needed in the real laboratory because all of the data has errors, and the real regression function is therefore not knowable.

3 Results and discussion

A summary of the results of the research to date suggests that information saturation could not occur until the sample set included the full range of all the independent variables, from minimums to maximums in the designed test program. In all of the successful demonstrations of information saturation, it can be seen that the RMS residual error FoM begins at a low level with the very first, while regressing the fewest samples, and then gets larger as the PSR matrix sample count increases. The very first PSR model dataset (PSR matrix) is for a sample count of one more than the number of unknown coefficients in the catalyst regression function. With so few unknowns and so few samples, a low error between model and data is expected, especially when using accepted first principles to select the regressor terms [12]. There is a natural reaction to conclude that this low error produces the best model. This is not assuredly true, and such assumption leads to poor models. As more samples are added in the PSR process, the errors grow as the model misses more data points. Eventually, and at times with very low sample counts, information saturation is reached, and the learning zone samples are consumed by the growing PSR matrix and the PSR analysis enters the convergence plateau. That growth of the error leads to a better model is counterintuitive, unless information theory is applied to the principles of regression. Regression can optimally fit only those samples that are given to it. Information saturation occurs because each new sample contributes no new information, and the error reaches its plateau. If new data points were to be obtained outside the hyperspace used for creating the model, there will likely be a rise in the convergence plateau and different regression coefficients. In short, there would be a new model based upon a new set of hyperspace vertexes. On the other hand, if a user of the model entered independent variables that are outside the tested modeled hyperspace, they would be extrapolating beyond the verified limits of the model. PSR analysis helps to explain why extrapolation of any empirical model is discouraged by those who use models based on regression.

3.1 Vertexing

Vertexing is a multi-step process, it begins with finding the best candidate data samples of vertexes. This can be at test planning stage, before entering the laboratory, or it can be done on legacy data that was collected orthogonally per ISO 4409. Vertexes are the outer corners that define the ultimate limits of the tested hyperspace. Vertexes consist of all possible combinations of minimums and maximums of the independent variables: pressures, speeds, viscosities, displacements. All four are not variables in all cases, e.g., fixed displacement machines.

Regarding vertexing, there are 2^n possible vertexes in a given test, where n is the number of actual independent variables that are implemented in a test program. Arrangement of the data samples is necessary so that all the vertexes are the first samples in the genesis file. The application of strategies for the arrangement within the vertex data samples is required to avoid data crowding and variables that have not undergone actual change within the beginning small sample counts of PSR analysis. This was caused by data crowding and volatility of the A_o coefficient in fig.2. In a fixed displacement test, there are only three independent variables, so there are 2^3 vertexes. A simple binary truth table can quickly construct all eight combinations of minima and maxima. This is a three-dimensional hyperspace and can be visualized [8]. In a variable displacement test, there can be four independent variables and the hyperspace is four dimensional and has 16 vertexes, and so on.

Each vertex is a data sample. When the data is sequenced into the PSR matrix such that the vertexes are the first 2^n points in the matrix, the resulting PSR analysis reaches information saturation at the conclusion of regressing those samples. The efficacy of starting PSR analysis with the vertexes is every bit as good as LHC sequencing. In fact, the most significant conclusion of this paper is that the optimal dataset for PSR analysis will have 2^n vertex samples followed by an LHC sequence with 2^n samples to form a super efficacious learning zone.

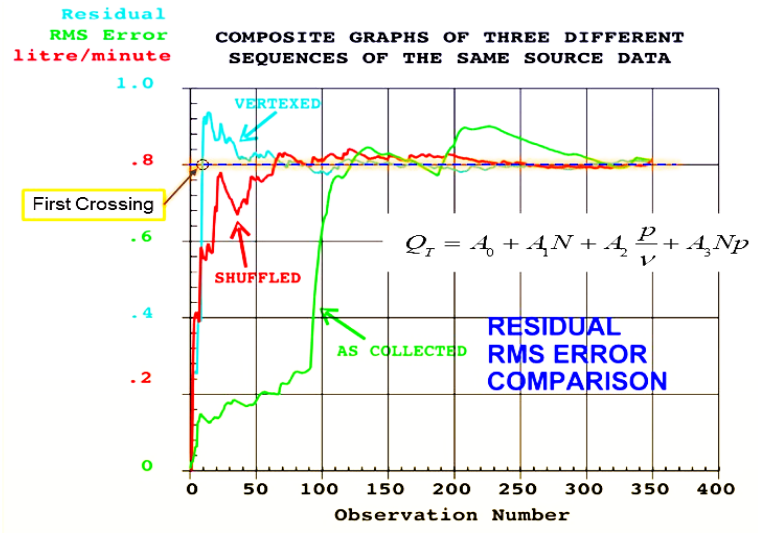

Figure 7: Overlay of compared data sequences.

The vertexes assure that all extremities of the hyperspace will be reached and represented while the LHC samples will assure that the interior of the hyperspace is considered. In the case of a two-dimensional test program, with only pressure and speed as variables, there are only four vertexes. It is recommended that the LHC sample set contain at least 8 samples in the case of a two dimensional hyperspace. Figure 7 is an example of how quickly the information saturation takes place with eight vertexes leading the genesis file to PSR analysis. The attainment of information saturation is identified as the point of "first crossing" in the figure. The resolution of the graph with 350+ data samples cannot display the gap between the vertical axis and the start of the two curves with a catalyst regression function with only four unknowns. That catalyst function was one that the authors examined as part of the many trial-and-error analyses during the early months of the project.

The authors' software for vertex mining has a min or max strategy as options. The min strategy arranges the vertex samples so that there is the smallest amount of hyperspace distance in going from one vertex to the next. The max option is just the opposite. It sets the order of adjacent samples so that the greatest amount of hyperspace distance separates vertex neighbors in the genesis file. There has not been enough exploration of these options and results to draw any conclusions, however, the authors lean toward the max as being preferred. It is known that both data crowding and non-varying parameters can occur within the vertex learning zone. By rearranging the vertexes so that there is a maximum change from one sample to the next, it is reasoned that both data crowding and non-varying parameters are broken up.

Figure 8 shows that the A_0 coefficient is more volatile than the RMS Residual error. This is to be expected, and is the reason that coefficients are not good indicators of the success or not of reaching information saturation. However, the peak-to-peak volatility in A_0 is small, ranging from about -0.5 to about -1.05 a spread of about 1

litre/minute. Given that the maximum measured flow from this pump was nearly 100 litre/minute, the volatility of A_0 is only about ±0.5% of maximum measured value. The second issue examines the final value of the RMS Residual error. It is nominally about 0.8 litre/minute, or about 0.8% of maximum measured flow. In comparing the RMS value to the graphs of fig. 2 the terminal error was only 0.38 litre/minute. Both curves began with the same source data, however, the catalyst regression function for fig. 2 had seven unknown coefficients, while the function for fig. 8 used only four unknown coefficients. It is well known among modelers of physical processes that the number of terms, that is, function complexity, will affect the data fit. Figure 3 displays the better fit of the two because the RMS Residual error is less.

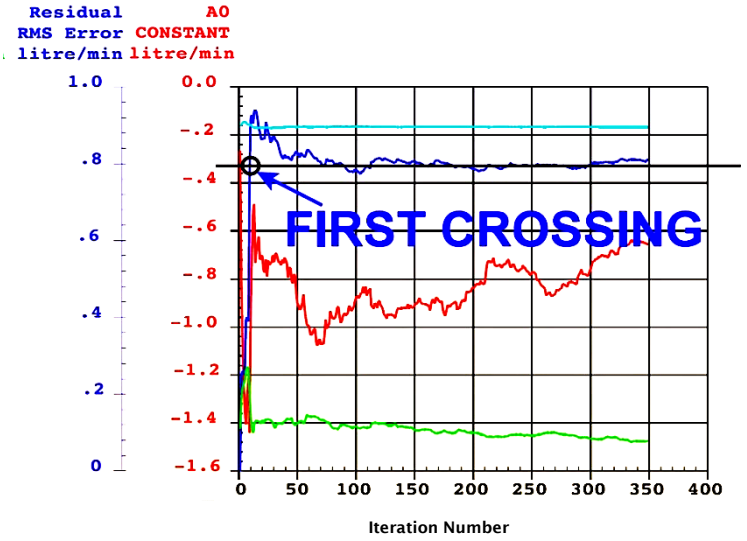

Figure 8: PSR Analysis results for an eight vertex sample dataset.

Vertex mining can be used on orthogonally collected data after the fact. It was used on legacy data for this paper. Several steps are required, and some special programming is needed in order to make the correct calculation of hyperspace distance. The steps are:

1) Set up a binary truth table listing all combinations of minimums and maximums based on the number of independent variables in the test program, i.e., 4, 8, 16, 32, etc. total combinations. 2) Search the existing data file for the absolute mins and maxes for each one of the independent variables in all samples of the data and save it in its own matrix. 3) Normalize each independent variable so that its maximum is 1.0 and all other values are less. 4) Normalize all corresponding values in the binary table so that each max value is 1.0 but each normalized minimum is whatever the normalization produces. 5) Taking each combination from the binary table, compare its proximity to each one of the data points in the total data array and save the one with the smallest hyperspace hypotenuse as the real vertex in a totally new matrix. Place that value at the front of the data array, but use the un-normalized original values. Then move to the next binary table entry and repeat the search for closest proximity. 6) When all binary table entries are completed, move all of the remaining samples into the new vertexed array making sure that the vertexes are not included because they have already been placed at the front of the new vertexed matrix.

From Johnson's progress reports [10]-[12], this phenomenon was seen clearly when the laboratory test sequence used LHC sampling strategy which also reliably and repeatably predicted the number of samples needed to reach information saturation. The results of shuffling the data to break up data crowding and non-varying independent variables, produced similar results, that is, information saturation was demonstrated, using the same data that failed in its original orthogonal order. Shuffling, however, had no predictor indications about how many samples were needed for information saturation. Genesis files for a gear pump and variable displacement piston pump were both shuffled. The PSR results are subject to interpretation, but the gear pump required about 30 to 40 samples to reach information saturation while the piston pump needed about double that number of samples. In another analysis of the piston pump by the authors, data was subjected to simple sorting on pressure. That is, the lowest pressure samples were first in the genesis file and progressively grew to maximum pressure with the last sample. The PSR results showed no clear information saturation. Instead, the RMS residual error continued to rise to the very end of the total 353 samples. This observation led to the conclusion that the learning zone needed to be populated with samples that had representation from the extremes of the independent variables. This was supported by the

demonstrated reality that LHC sampling forced the test points to range over the extremes of the independent variables. The natural outcome was the postulate that the best representatives for the learning zone would be the vertexes of the actual tested hyperspace. A variable displacement pump tested at two or more viscosities and several displacements has four independent variables, and therefore 16 vertexes are needed to define the boundaries of its hyperspace, while a gear pump, with its fixed displacement, would have only 8 vertexes in three dimensions.

A case study was carried out with a simple, and non-algorithmic, test sequence starting with the eight vertexes of a piston pump test. The goal of this case study investigated if sampling the vertexes of the hyperspace alone formed the learning zone with a consistent minimum sampling size. For example, a sample size of 16 vertexes for a four-variable test and eight vertexes are required for a three-variable test. Because the data from the pump was already obtained before the case study additional testing was not convenient, so a peripheral program was written to harvest the vertexes from the existing orthogonal legacy test data by applying a basic data mining strategy. A learning zone consisting of the vertexes of the test program hyperspace would assure that the full ranges of all independent variables would be present in the learning zone, fig. 7. A new test program with a learning zone comprised of only the hyperspace vertexes was not practical at the time, thus the synthetic process of harvesting vertexes from existing legacy orthogonally sequenced data was the only timely approach, under the circumstances. The mining and harvesting involved scanning the source data file for the extremes of all independent variables that were used to formulate an ideal set of vertexes. Next, the ideal vertexes were used to locate the real data samples that were closest to the ideal vertexes. In the first use of the method, there were three independent variables, requiring eight vertexes. When the nearest samples were found, they were extracted from the original genesis data set and reattached, but as the opening eight samples. Performing PSR analysis on the modified (re-sequenced) data set produced efficacy that rivaled the results of LHC sample sequencing. Because they represent the samples with the greatest spread in data, the eight mined vertexes were used as a learning zone in order to test the idea that the vertexes will be as efficacious as is produced by LHC sampling. The vertex data sequence has been referred to as a "synthetic" learning zone, but only because the first efficacious learning zone used LHC sampling in a more or less conventional design of experiment process. "Synthetic" in this context, suggests that the vertexes were not part of a pre-laboratory designed test plan.

Given that the StartSample variable was set to 5 (refer to the 4-unknown catalyst regression function displayed in Fig. 7 and Fig. 8) which makes the actual number of regressed samples at the first crossing point about 10 or 13, depending on the readers' interpretation of the graph, it is unambiguous evidence that indeed, vertexes produce learning zone efficacy that is similar to, if not better than that of LHC sampling. Perhaps a combination of the two could produce even better results. The opportunities to improve efficacy have only been explored to a very limited degree. This has led the authors to postulate that a more optimal designed test sequence will use both LHC and vertexes in the learning zone. Other variables are typical of what would be expected with the efficacy of LHC sampling. The A_0 coefficient in fig. 8 has some opening volatility, but it's not excessive and settles down reasonably well. Beyond the learning zone, the RMS peak-to-peak value is about 0.6 lpm, and with a maximum measured output of about 100 lpm, it calculates to about 0.5% of maximum output total spread (±0.25%), which is only half of the ±0.5% uncertainty limit required for the output flow per ISO 4409 (ISO, 2019) when claiming Class A Measurement uncertainty. To this point all the presented individual PSR analysis results are from a data file that was originally collected using classical orthogonal data sequencing. This is the method suggested by the recommendations of standard method ISO 4409 [14]. The results are shown in graphical form in fig. 8.

It is well known that the results of classical regression, that is, using all samples to produce a single model, will always be identical when using the same data, but in various sequences. That is, the terminal, or final coefficients, and FoMs will be exactly the same, regardless of data order. This truth is borne out in figs. 5, 6 and 7 in the fact that all traces converge on the same final value for the Residual RMS Error regardless of data order, however, the trajectories taken to get to the plateau depends on the data order. The data can be rearranged manually, especially at the beginning of the PSR analysis, so that information saturation is reached at a lower sample count. Two reliable and predictable means to achieve information saturation are now known, vertexing and LHC sampling.

4 Conclusions

Research results to date demonstrate that PSR analyses are affected by the order in which data is presented to PSR's iterative regression process. In setting up a genesis file, it is the skill with which the samples are arranged that control whether or not information saturation is demonstrated more so than the numbers of samples. The same research has identified three different means for establishing efficacious sample sequencing that are objective and

predictive: The first is through the use of LHC experiment design to create the pre-laboratory test program for both the learning zone and the convergence zone. The second is to include the vertexes of the hyperspace extremities to guide the collection of first data samples, and the third is to randomly rearrange legacy data that was collected using the conventional orthogonal sequencing that is recommended in ISO 4409. The first two of these methods will assure that information saturation is reached within the number of samples requested of the LHC algorithm and the number of vertexes needed to define the test hyperspace. Random rearrangement of legacy orthogonal data shows promise of assuring information saturation, however, it is less predictive in the number of samples required. A fourth method, sample randomization by shuffling, produce information saturation, however, it was not predictive and was abandoned. Incorrect data sequencing (sample order) has a negative effect on the results. False negatives and false positives can occur in PSR Analysis, as illustrated in Fig. 2 and Fig. 3. The PSR analysis paradox is that one needs to have many samples in order to conclude that fewer samples would have sufficed. This makes PSR analysis a research tool and a means of verifying the efficacy of a given learning zone strategy to reach and display information saturation. PSR analysis will continue to be a useful tool in mathematical model development as a means for evaluating and validating conclusions about model quality.

5 Future Research

Future research efforts are aimed at finding a means for converging on the actual number of samples needed to achieve information saturation in the learning zone and to improving the information that can be gleaned from a data set, aside from the quest for sample size. It is postulated that a combination of LHC and vertexed learning zones will give test labs the tool to predict minimum sample count for producing reliable and useful math models. This theory needs additional testing that consists of designing the test programs with LHC and vertexing combined into a single learning zone. There needs to be one more round of testing to evaluate the efficacy and expected optimality of a combined LHC and vertexed learning zone. All research to date regarding PSR analysis has not given a means for positively calculating the minimum number of samples needed to achieve a useful model. The next step in this continuing research program aims to correct that shortcoming. The new research must investigate the nature of the learning zone data samples by comparing the sequencing that produces information saturation in the learning zone to those which don't devise a discrimination theory and then test it. This could lead to a practical Efficacy Index. At the same time, the models themselves, that are taken from the limited data sets, must be compared to ensure that they produce accurate projections of such quantities as output flow and input torque as well as overall efficiency.

6 Acknowledgments

The authors wish to thank Mr. John Montague for providing his comments and editorial revisions to this article.

7 Nomenclature

Designation	Denotation	Unit
F_i	Force	N
p	Pressure	Pa
LHC	Latin Hypercube	
PSR	Progressively Sequenced Regression	
ν	Kinematic viscosity	
N	Rotational shaft speed	rpm
p	Pressure differential	
A_n	Regression coefficients	
q	Volumetric flow rate	lpm

8 References

[1] Love, L., Lanke, E., Alles, P. (2012), "Estimating the Impact (Energy, Emissions and Economics) of the U.S. Fluid Power Industry," Oak Ridge National Laboratory: 2012, Report No. ORNL/TM-2011/14.

[2] International Organization of Standardization ISO, (2008). "ISO 8426 - Hydraulic fluid power — Positive displacement pumps and motors — Determination of derived capacity." ISO copyright office.

[3] Society of Automotive Engineers (2019), "Hydraulic Power Pump Test Procedure J745_201911." DOI: https://doi.org/10.4271/J745_201911

[4] Wilson, W. E. (1950). Positive-displacement pumps and fluid motors. Pitman Publishing Corporation.

[5] Toet, G. Die Bestimmung des theoretischen Hubvolumens von hydrostatischen Verdrangerpumpen und Motoren aus volumetrischen Messungen. Olhydraulik Pneum. 1970, 14, 185–190.

[6] Eggers, B., Rahmfeld, R., & Ivantysynova, M. (2005). An energetic comparison between valveless and valve controlled active vibration damping for off-road vehicles. In Proceedings of the JFPS International Symposium on Fluid Power (Vol. 2005, No. 6, pp. 275-283). The Japan Fluid Power System Society.

[7] Conrad, F., Trostmann, E., & Zhang, M. (1993). Experimental identification and modelling of flow and torque losses in gerotor hydraulic motors. In Proceedings of the JFPS International Symposium on Fluid Power (Vol. 1993, No. 2, pp. 677-682). The Japan Fluid Power System Society.

[8] Johnson, J.L., (2018a). "Design of experiments and progressively sequenced regression are combined to achieve minimum data sample size", Int. J. Hydromechatronics, Vol. 1, No. 3, pp.308–331.

[9] McKay, M. D., Beckman, R. J., & Conover, W. J., (1979). "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code." Technometrics, 21(2), 239-245.

[10] Johnson, J. L., (2018b). "Progressively Sequenced Regression Helps to Establish Minimum Sample Size at Test Time." Unpublished report, 1 of 4, distributed to active members ISO\TC131\SC8\WG13, Mathematical Modeling Ad Hoc Project, January, 2018.

[11] Johnson, J. L., (2018c). "A Sequel - Progressively Sequenced Regression Helps to Establish Minimum Sample Size at Test Time." Unpublished report, 2 of 4, distributed to active members ISO\TC131\SC8\WG13 Mathematical Modeling Ad Hoc Project, February, 2018.

[12] Johnson, J. L., (2018d). "Progressively Sequenced Regression Is Useful When Coupled with Efficacious Set Point Randomization." Unpublished report, 3 of 4, distributed to active members ISO\TC131\SC8\WG13 Mathematical Modeling Ad Hoc Project, March, 2018.

[13] Panwar, P., and Michael, P., (2018). "Empirical modelling of hydraulic pumps and motors based upon the Latin hypercube sampling method." Int. J. Hydromechatronics, Vol. 1, No. 3, pp. 272–292.

[14] International Organization of Standardization ISO, (2019). "ISO 4409 - Hydraulic fluid power - Positive displacement pumps, motors and integral transmissions - Methods of testing and presenting basic steady state performance data." ISO copyright office.

[15] Viana, F., Venter, G., Balabanov, V., (2010). "An Algorithm for Fast Optimal Latin Hypercube Design of Experiments." International Journal for Numerical Methods in Engineering; 2010.

[16] Helton, J.C., and Davis F.J., (2002). "Latin Hypercube Sampling and Propagation of Uncertainty in Analyses of Complex Systems." Sandia Report, SAND2001-0417, Sandia National Laboratories, New Mexico (November 2002).

[17] Zubarev, D. I., (2009). "Pros and Cons of Applying Proxy-models as a Substitute for Full Reservoir Simulations." Society of Petroleum Engineers. doi:10.2118/124815-MS.

[18] Vose, D., (2014). “The pros and cons of Latin Hypercube sampling.” Vose LinkedIn channel. Accessed 2018, url: https://www.linkedin.com/pulse/20140708131747-483951-the-pros-and-cons-of-latin-hypercube-sampling.