

# COMPARISON WITHIN AN ENVIRONMENT

# INDEX

## **1- intro**

## **2- preliminary ideas**

2.1- definitions

2.2- overview

## **3- local environment**

3.1- definitions

3.2- overview

3.3- defining an environment

## **4- direct relationships**

4.1- intro

4.2- multivariate problem

## **5- comparisons**

5.1- from local to global understanding

5.2- system of reference

5.3- wide-sense understanding

5.4- learn-understand → predict

## **6- conclusions**

## **appendix I - negative numbers**

## 1- intro

“comparison within an environment” represents the formal mathematical background sustaining a new form of information maximisation [proper understanding → reliable forecasting] - because of the huge amount of resulting calculations [detailed combinatorics & trial-and-error process based] and the subsequent reliance on highly-specialised software, no specific numerical developments will be included in this paper

## 2- preliminary ideas

some concepts have to be clearly delimited before continuing

NOTE - by default, any **behaviour/environment** reference will be done from a **local point of view**  
NOTE 2 - local/global distinction will be explained in section 5

### 2.1- definitions

**behaviour** – any situation where the following elements can be identified

1. actuator → someone or something provoking the action
2. action → any **measurable** effect, consequence, implication, etc.

note that, in a wide sense, almost everything might be somehow defined as a behaviour

**measurable** – any variable suitable to be included within some specific scale [virtually, anything]

examples of arbitrary scales

- woman, man, kid → 1,2,3 or 2,3,1 or 3,1,2... – any ordering is acceptable
- yellow, orange, red → 1,2,3 or 3,2,1 – only configurations respecting its inherent scale

**defining a behaviour** – two steps

1. actuator definition [independent variables] → combination of all the phenomena having any effect over the corresponding action  
virtually infinite configurations - practically, less than 10 variables can define, within an acceptable error level, the most of the natural situations
2. action definition [dependent variable(s)] → measured [or estimated] effect

**human vs. mathematical understanding** –

- a. ideas/concepts/sentences → equations
- b. understanding/learning process → forecasting methods [regression analysis]

NOTE - from this point onwards, “understanding” will refer to mathematical understanding

**pure understanding** – no external source takes part within the understanding process [just the raw data] - user intervention limited to mere “interpreter” [actual applicability of the result, is the training set representative enough? (global understanding or overfitting?), etc.]

**artificially-extended understanding** – additionally to raw data, some indications are given by an external source [user or automated system] - these inclusions try to overcome restrictions of the given understanding proceeding [usually, maximum number of independent variables]

**artificial extension** – many theories have developed means to overcome the restrictive configuration of the ideal understanding methodology [regression analysis => maximum of 2 independent variables]

main types [examples]

1. user-defined parameters [time series] → more-or-less directly related to the main problem, although imply an excessive increase in the level of uncertainty and a difficult-to-justify partial renounce to the (understanding-)decision-making [user intervention quite influential over the final result]
2. conditioned measurements [DOE] → in order to accept as many independent variables as possible, some methods perform multiple-step simple regressions - at each step, all the independent variables are kept constant except one [the only one taking part in the calculations] - this configuration involves many combinations/calculations [values\_per\_variable<sup>number\_of\_variables</sup>] what provokes that only specific values [usually, just 2 per variable (maximum & minimum)] are considered as system-feed and thus specific measurements have to be taken
3. indirect solutions [PLS regression] → models created in base of some statistical variables offering a description for the given behaviour with the sole utility of peer-to-peer comparison [more, less, higher, better, worse... than...], never giving to-the-point solutions

## 2.2- overview

almost any situation/idea/behaviour... can be translated into understandable numerics, that is, located within some [natural or completely arbitrary] scale

the lack of a complex-concepts-based-understanding [or human understanding] system in actual mathematics drives any method to start from quite simple principles: iterative calculations and trial-and-error-based systems

the ideal methodology [regression analysis] has been formally developed just up to 3 dimensions [too simple configuration for the most of the situations] - the solution given to this problem will condition drastically the accuracy and applicability of any global-understanding procedure

### 3- local environment

no absolute understanding is possible [some system of reference is always required] and this is precisely the reason explaining the introduction of this new concept [“relativising” mathematical understanding]

#### 3.1- definitions

**boundaries** – upper and lower values for any variable [from the training set]

**variations** – mathematical evolutions [between both boundaries] experienced by the values of every variable describing [the actuator/action of] a behaviour

**collateral variations** – variations directly provoked by changes in other variables describing the given behaviour - example

$$\text{dep\_var}_1 = \begin{cases} 3 \cdot \text{indep\_var}_2 & , \text{ if } \text{indep\_var}_3 \in (-\infty, 5] \\ 2 \cdot \text{indep\_var}_2 & , \text{ if } \text{indep\_var}_3 \in (5, \infty) \end{cases} \quad \text{where,}$$

dep\_var<sub>1</sub> → action description  
indep\_var<sub>2</sub> & indep\_var<sub>3</sub> → actuator description

**multi-behaviour** – mathematical understanding accounting for collateral variations - various environmental-conditions-dependent equations are calculated

#### 3.2- overview

local environment can be defined as “the whole picture” for each couple actuator-action [case or training point], as the combined evolutions of all the involved variables [including their respective inter-relationships]

more specifically [and always from a local point of view]

behaviour – local understanding, usually describable by attending just at certain variables

environment – ideal framework allowing to calculate this behaviour - all the variables [the ones being accounted for the corresponding analysis] have to be considered

#### 3.3- defining an environment

no general rules can [and, in any case, should] be fixed for a so important [and specific-conditions-dependent] process, just a experienced-enough voice in the corresponding field could provide some relevant insight

the recommended [but fully adaptable] steps to be taken are

1. isolate the action [dependent variable]
2. determine what is the group of phenomena offering an actuator’s better description [independent variables]

### 3. boundaries

- a. description [independent variables] → max. & min. values within the given behaviour
- b. action [dependent variable] → values for the corresponding max. & min. (extreme) actions resulting from the given actuator description - example

independent variables [actuator description]	{	<b>A</b>	1	4	2
		<b>B</b>	2	1	9
dependent variable [action]	{	<b>C</b>	3	5	11

by assuming [it is impossible to conclude anything by attending just at 3 points] that this behaviour is properly described with the equation  $C=A+B$ , the boundaries for C would be 2 and 13

## 4- direct relationships

despite representing one of the simplest mathematical methodologies, it constitutes the ideal first step towards complex-behaviour understanding

### 4.1- intro

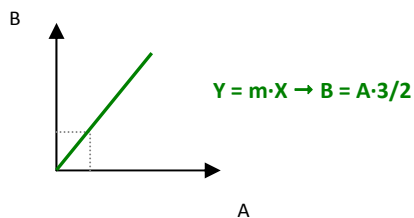
anyone's school basic learning is

$$A = 2 \rightarrow B = 3$$

$$A = 4 \rightarrow B = ?$$

if 2 implies 3, how much would 4 imply?  $\rightarrow B_2 = 3 \cdot 4 / 2 = 6$

what is really going on here?



by assuming some sensible reference point [ $A=0 \rightarrow B=0$ ], a straight line has been drawn [and thus a regression performed]

after applying some of the previously introduced ideas

actuator description  $\rightarrow A$  [IF  $A = \dots$ ]

action  $\rightarrow B$  [THEN  $B = \dots$ ]

the aforementioned relationship represents the most perfect mathematical understanding of the given problem, although it is highly unreliable [just two points, out of them one is an assumption]

### 4.2- multivariate problem

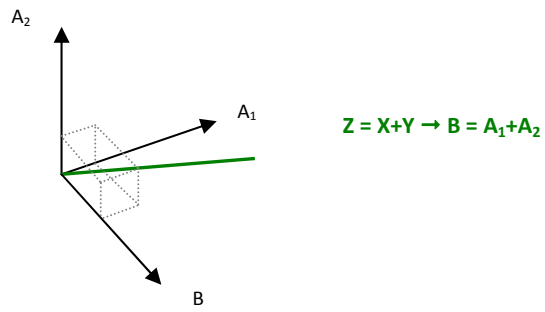
would it be possible to solve more complex problems by applying direct relationships? problems involving a higher number of [independent] variables type A?

for example

$$A_1 = 2 \ \& \ A_2 = 3 \rightarrow B = 5$$

$$A_1 = 3 \ \& \ A_2 = 5 \rightarrow B = ?$$

note that all the given values are varying at the same time and thus the only valid proceeding is finding some 3-D fit [2 independent vs. 1 dependent]



actuator description  $\rightarrow A_1$  &  $A_2$  [IF  $A_1 = \dots$  AND  $A_2 = \dots$ ]

action  $\rightarrow B$  [THEN  $B = \dots$ ]

note that conventional regression methods are applicable just for problems involving up to 2 independent variables, from this point onwards artificial extensions have(?) to be applied

## 5- comparisons

local environment [each case/training point] → local/absolute/isolated... behaviour

comparison → generalisation/relativisation/joint... of various local behaviours [in order to UNDERSTAND them]

global environment [combination of all the cases/training points under consideration] → (relative) system of reference resulting from the comparison among [the understanding of] local environments

### 5.1- from local to global understanding

once the environment has been built [e.g., fuel consumed by an engine as a function of seven operating variables] and further cases being located [e.g., datasets from 30 different cars] the next step is actually understanding such a behaviour as something global, converting the potentiality into reality, the starting point and the end point into actual movement

can the fuel consumption be calculated as a function of the given variables?, does this behaviour actually exist? the transition local to global answers this question - setting the local environment [by selecting these 7 variables after just considering one car] was just a preliminary assumption, confirming [or rejecting] its validity [by comparing this local environment against 30 additional ones] is precisely (globally) understanding - the global environment has been created and a new (relative) system of reference appeared

### 5.2- system of reference

nothing can be understood without being referred to certain system [allowing the conversion absolute into relative]

NOTE - local environments represent the system of reference for local understanding, although this "understanding" is no more than a "transitional mathematical trick" [although necessary], something allowing to reach the global environment, the real system of reference for the real [global] understanding

the most common proceeding is building some absolute system of reference, a global-and-solid-enough ultimate truth sustaining the whole system [e.g., constant speed of light for Einstein's theory] but why introducing a so restricting condition? why conditioning the validity of all the ideas to something certain today but perhaps invalid tomorrow?

the comparisons themselves constitute a (**relative**) system of reference, a global environment - how good is this behaviour/local environment? compare it against the rest of the available ones [global environment]

### 5.3- wide-sense understanding

up to this point all the references to (mathematical) understanding were done "on a wide sense", as general term including learning/training [numerics] and strictly-speaking understanding [numerics + (common-sense-based) delimiting rules]

training/learning represents the pure-mathematics part - some regressions [or equivalent methodologies] are performed over the given dataset [training points] looking for the ideal fit explaining the underlying behaviour

the understanding part represents a look-beyond numbers warning - a set of adaptable-enough rules has to be developed in order to restrict as much as possible any possible arbitrariness

#### 5.4- learn-understand → predict

direct-relationships, as introduced in the previous section, can provide a global and clear enough picture about mathematical understanding without relying on detailed mathematical developments, which could provoke some misunderstandings

regressions [or any “artificially-extended” method] represent the way to mathematically understand [IF... THEN] and correspondingly to forecast [IF... THEN (in the past), IF (in the future)... **most probably THEN**]

main milestones within the whole process

- a. training [learning] → past reliable datasets [regressions]
- b. understanding → set of sensible conditions [common sense before numbers]
- c. forecasting → new predictable datasets [application of calculated fits]

NOTE - this is just a general theory providing some guidelines describing a different way to face data analysis, not a manual - the confidence spans, defining values, specific boundaries, etc. should be developed by experts in the different fields of expertise [overall-applicable rules => errors]

##### **a. training**

regression analysis [or any equivalent method] is being carried out over the given dataset

NOTE - although no specific indication about the type of method to be used [or how the extension-beyond-2-independent variables problem has to be sorted out] will be provided in order to build a generic and adaptable enough theoretical framework, it is straightforward that any methodology applying this theory should be based on combinatorics and trial-and-error analysis

additionally to the most important issue here [what understanding method (type of regression or equivalent methodology) to use], some minimum conditions have to be fixed - for example: historical pre-study [making sure that the evolution is more-or-less stable], not less than 30 training points [any behaviour defined by a lower number would be consider as pure casualty], looking for a set providing wide-enough boundaries [making sure that will wrap all the points to be forecasted and hence no extrapolation will be carried out] , etc.

##### **b. understanding**

mathematics are dumb by default - some common-sense reliance is much more consistent than any “numbers say so”

hence an open and adaptable enough structure is the only acceptable starting point for any system aiming to understand - again a set of conditions capable of distinguishing between “good numerics” and “proper understanding” has to be established: rules avoiding overfitting [e.g., 10 training points or more in order to consider third-degree polynomial fits], setting confidence spans [although an automatic determination of this variable can be really dangerous], determining what it the best

understanding [virtually infinite fits can be calculated from a medium-difficulty set (above 5 independent variables), this sub-system will create a “scale of goodness”], etc.

### **c. forecasting**

a new set of conditions has to be created: never extrapolating, despite confidence spans or error levels within the training set [understanding] fixing some upper in-the-safest side limit, setting some “emergency” sub-system [as soon as any case shows an error beyond X = predictive capabilities assumed to have expired + new training is required ]

## 6- conclusions

comparison within an environment represents an open and adaptable enough framework providing some general guidelines to face data analysis [understanding → forecasting] by relying on the following ideas

- detailed combinatorics + iterative learning [trial-and-error process]
- regression analysis [direct relationships] based
- relative system of reference [scale of goodness giving by your current knowledge]
- a clear [and excluding] bipartition has to be respected
  1. maths → method training [finding out the most describable equations]
  2. user → common sense to maths' conclusions, final decision-maker [delimiting rules]

as less user intervention as possible in part 1 & no mathematical reliance on part 2

## appendix I - negative numbers

some methods could find problems to naturally deal with negative numbers [consequence from restrictions aiming to allow in-the-safest-side combinatorics] - a good solution [having present that the whole methodology is based on comparing variations] would be developing a “re-scaling” subroutine

re-scaling means moving the internal reference point for the calculations [from  $-\infty$ ] “towards positivity” in order to avoid negative values [by converting them into positive ones] - usual proceeding: adding some [high enough] number to ALL the values of all the [involved\*] variables

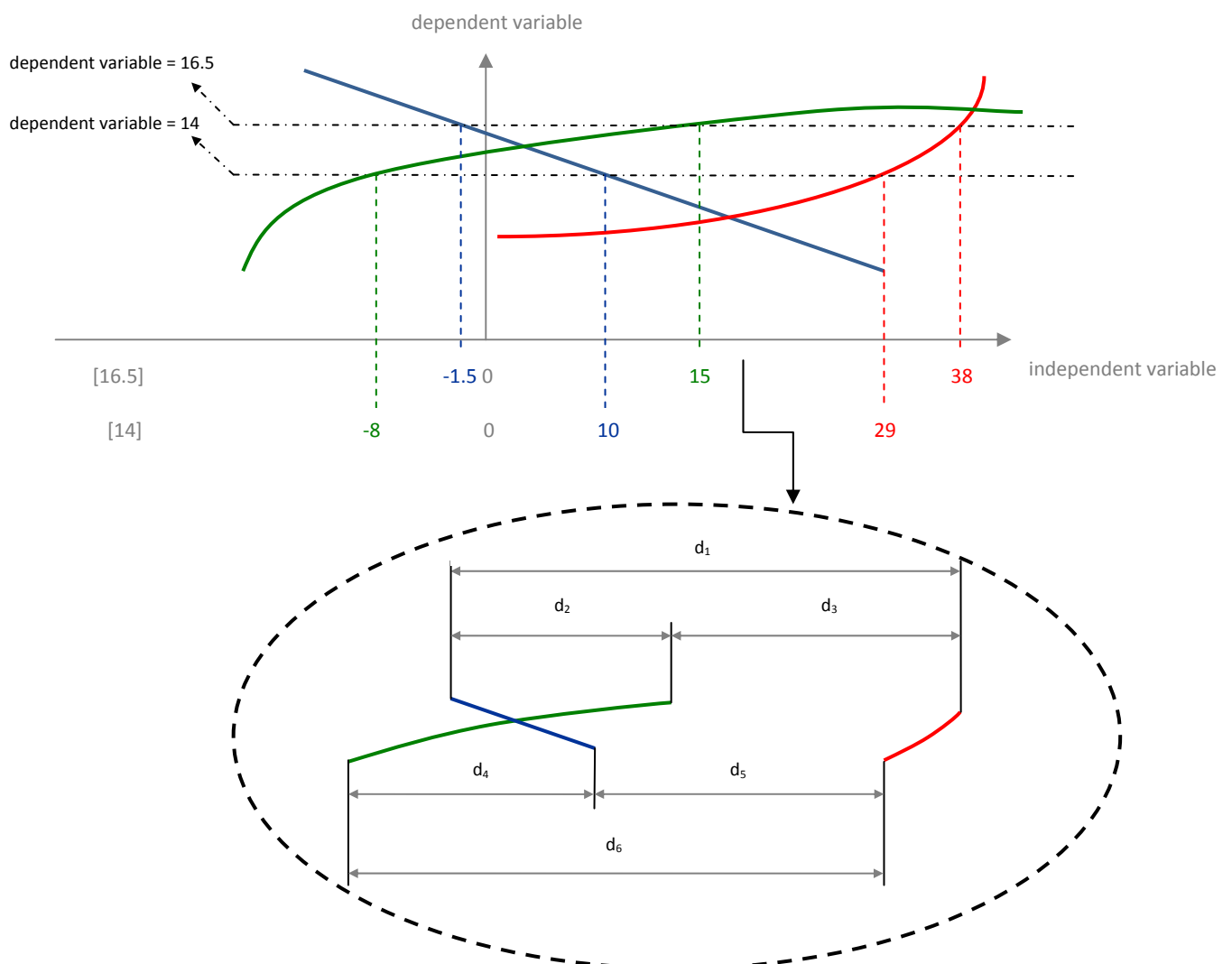
\* by assuming only negative values in X-axis [graphs below], Y-values have to be kept constant



### EXAMPLE 1 –

3 independent variables affecting a dependent one

below a 2-D graph showing the evolutions independent/dependent variables and the values adopted by the first ones for two different positions of the last one

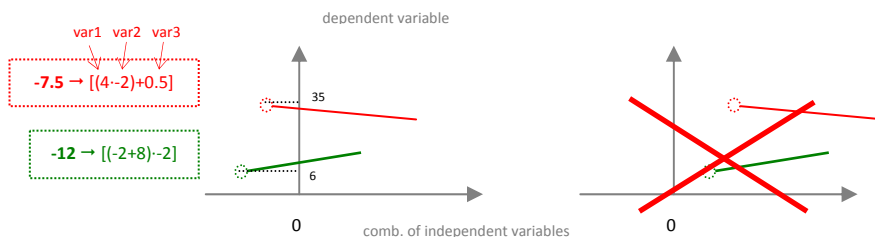


above, in the highlighted area, the most relevant relationship [relative distances among the values taken by each variable under the same independent “level”] has been included → independently upon the calculation method being used, these “distances” have to remain unaltered in order to allow a proper comparison, and hence a proper understanding

**is it possible to perform “this movement” in any case?**

in the above example, it is completely irrelevant how long the moving distances become: as far as the values for the independent variable remain unaltered, the comparison will not be affected - however, depending upon how the corresponding calculation method deals with “the multivariate problem”, this issue might become a big concern

in the 2-D graph below these lines, the values of the dependent variable are plotted against the ones resulting from a combination of the independent variables



anyone can see that the proposed “movement towards positivity” is much more complicated to be performed - as explained above, the same value has to be added to EACH [independent] variable in order to keep constant the corresponding [relative] distances allowing the comparisons

giving a general enough and fully adaptable solution to this problem is virtually impossible [or, at least, it would required from a quite complex sub-algorithm dealing with it]

a more casual approach seems to be the ideal proceeding to sort this issue out, but this is something every calculation algorithm will have specifically to care about