Analysis of Different Curved Regressions using Free Software and Selection of the Appropriate Type Based on Statistical Tests for Goodness of Fit and Analysis of Variance

R.J. OOSTERBAAN

Retired from International Institute for Land Reclamation and Improvement (ILRI)
Wageningen, THE NETHERLANDS

https://www.waterlog.info

Abstract: There are various kinds of curved (non-linear) regressions, for example those based on the power function, the second degree polynomial (i.e. the quadratic function), the S-curve functions, and the third degree polynomial (i.e. the cubic function). Free software to apply these functions is available as SegRegA. It uses a generalization of the standard functions by applying a transformation the values of the independent variable before fitting the selected function. Under the conditions at hand, one the these four kinds might be preferable. The decision can be made with the R^2 magnitude, being the coefficient of explanation, for goodness of fit. When needed, confidence limits of R^2 may be employed to tests its significance. The R^2 is not always the ultimate criterion. Analysis of variance may also be required as well as theoretical considerations. In this article, data from literature on temperature trend in time and crop production against soil salinity as well as against depth of the water table are given as examples.

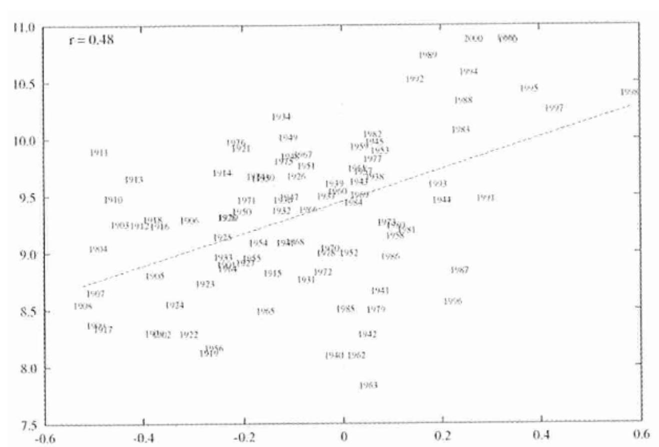
Key words: - Curved (non-linear) regression, free software, coefficient of explanation, goodness of fit, confidence belt, analysis of variance, statistical significance testing

1. Introduction

Curved, non-linear, regression may be tried when the linear relation between the dependent variable (Y) and the independent variable (X) shows a trend that deviates from a the straight line.

An example is shown in *figure 1*, published by the Dutch Meteorological Institute (KNMI) in de Bilt [*Ref. 1*]. A linear trend is assumed between the vertically and horizontally plotted data.

In the figure it can be seen that, at X values higher than 0.2 (i. e. in the right hand corner above), the majority of the Y values lies above the straight line. Here, the trend indicates a deviation from the straight line. One could ask whether a curved regression would give a better fit.



Vertical: annual average temperature in de Bilt, Netherlands
Horizontal: annual deviation of world average temperature

Figure 1. Screen print from the KNMI report. A linear trend is assumed between the vertically and horizontally plotted data The data plots are labeled with year numbers.

In the figure it can be seen that, at X values higher than 0.2 (i. e. in the right hand corner above), the majority of the Y values lies above the straight line. Here, the trend indicates a deviation from the straight line. One could ask whether a curved regression would give a better fit.

In another example, the relation between the yield of the potato variety 927 and the salt concentration of the soil has been observed to consist of an initial horizontal line followed by a descending line (*figure 2*), the breakpoint representing the crop salt tolerance as defined by the much used Maas-Hoffman model [*Ref. 2*].

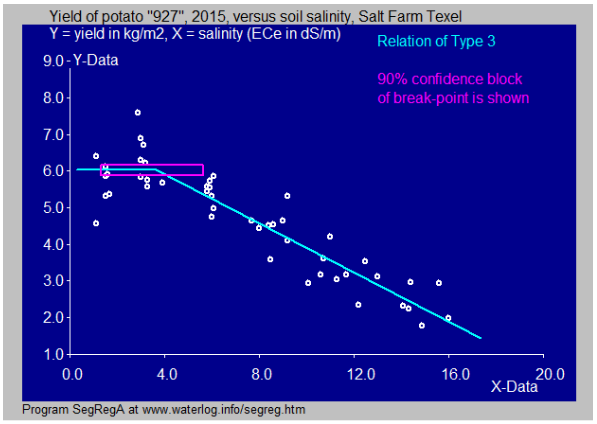


Figure 2. Relation between yield of potato variety 927 (Y) and soil salinity (X) according to the Maas-Hoffman model (Type 3).

In *figure 2* it appears that actually the trend of Y values for X values less than 4 is ascending rather than horizontal *(figure 3*).

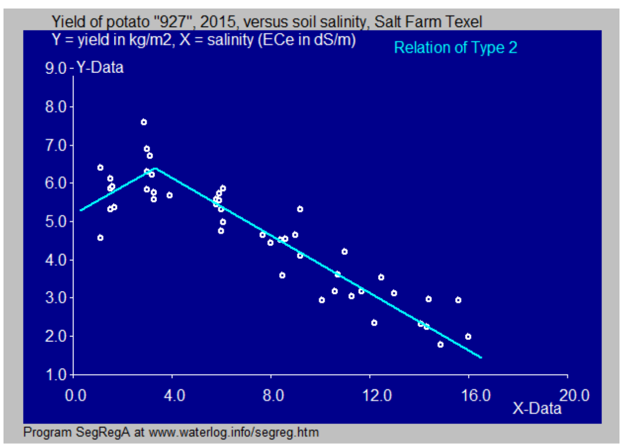


Figure 3. Relation between yield of potato variety 927 (Y) and soil salinity (X) using segmented regression Type 2.

Figures 2 and 3 were prepared using the SegRegA program [*Ref. 3*] employing the option of segmented regression types.

In continuation it will be investigated whether under such conditions standard or generalized curved regressions are feasible.

2. The principle of generalization of curved regression functions.

Curved regression functions can be of one of the following types:

1. Power function
2. Quadratic function
3. S-curve function
4. Cubic function.

To increase the applicability and goodness of fit, SegRegA [*Ref. 3*] uses X transformations with the help of the minimum value of X and an optimized exponent E before applying the standard curved function. This principle is called generalization. An overview is given in *Table 1*.

In the case of the S-curves [*Ref. 4a*, *Ref. 4b*], additional transformations are made to enable linear regressions for the determination of A and B (see *Table 2*).

The method of selection of the type of curved function in SegRegA is illustrated in the *Appendix I*.

Table 1. Generalized curved functions used in SegRegA, where:
Xmin = minimum X, Ymin=minimum Y, Ymax = maximum Y, W = X – Xmin,
E = exponent optimized by maximizing the goodness of fit to realize the generalization

Power function	$Y = A.W^E + B$ **#)**
Quadratic function, generalized	$Y = A.W^{2E} + B.W^E + C$
S-curve logistic, generalized S-curve Gumbel, generalized S-curve Gumbel mirrored (inverted)	$Y = (Ymax-Ymin)/[1+\exp(A.W^E+ B)] + Ymin$ $Y = (Ymax-Ymin)*\exp[-\exp(A.W^E+ B)] + Ymin$ $Y = (Ymax-Ymin)*\exp[1-\exp(A.W^E+ B)] + Ymin$
Cubic function, generalized	$Y = A.W^{3E} +B.W^{2E} + C.W^E + D$ **#)**

#) The regression with quadratic and cubic functions (also called 2nd and 3rd order polynomials or 2nd and 3rd degree polynomials) using the least squares method is done by matrix algebra and determinants as described in https://neutrium.net/mathematics/least-squares-fitting-of-a-polynomial/

Table 2. Transformations of generalized S-curve functions to determine parameters A and B by linear regression. They are based on the logistic [Ref. 5] *and* Gumbel *probability distributions* [Ref 6].

S-curve logistic	1st transformation: $Yt = (Ymax – Ymin)/(Y – Ymin) – 1$, so that: $Yt = \exp(A.W^E+ B)$ 2nd transformation $Yv = \ln Yt$, so that: $Yv = A. W^E + B$, so that: the parameters A and B can thus be found from a linear regression of Yv on W^E, while the exponent E has to be numerically optimized by Trials to effectuate the generalization.
S-curve Gumbel	1st transformation : $Yt = (Y – Ymin) / (Ymax – Ymin)$, so that: $Yt = \exp[–\exp(A.W^E+ B)]$ 2nd transformation: $Yv = \ln[– \ln(Yt)]$, so that: $Yv = A. W^E + B$, so that: the parameters A and B can be found from a linear regression of Yv on W^E, while the exponent E has to be numerically optimized by trials to effectuate the generalization.
S-curve Gumbel mirrored (inverted)	1st transformation : $Yt = (Y – Ymin) / (Ymax – Ymin)$, so that: $Yt = \exp[– \exp(A.W^E+ B)]$ 2nd transformation: $Yv = \ln[1– \ln(Yt)]$, so that: $Yv = A. W^E + B$, so that: the parameters A and B can be found from a linear regression of Yv on W^E, while the exponent E has to be numerically optimized by trials to effectuate the generalization.

Summarizing the principle of the generalized functions, it can be said that, before applying the standard functions know from literature, the X-values are transformed to a value $W^E = (X – Xmin)^E$ using a minimum value Xmin of X and an exponent E. Both are to be found by assuming a large range of values and selecting those values that lead to the best goodness of fit of the observed and simulated Y values.

Note that, when the exponent E is not equal to 1, the quadratic and cubic function are in fact not anymore exactly quadratic or cubic.

The aim of this paper is to demonstrate the possibilities of the use of generalized curved regressions and the considerations that can be followed to decide whether they are applicable in a certain situation and, if so, how to decide which one would be preferable.

3. Analyzing the average temperature trend in De Bilt, The Netherlands (1901 – 2020)

The SegRegA results for the four curved regression types are shown in respectively figure 4, 5, 6 and 7.

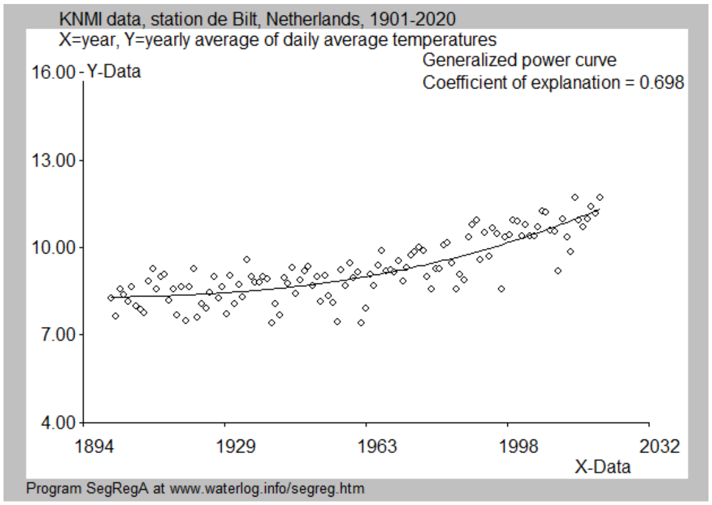


Figure 4. Temperature trend with fitted power curve:
$Y = 5.04\ (X - 1882)^{2.70} + 8.29$ *with* $R^2 = 0.698$

As an example, the SegRegA output file of the power curve case is demonstrated in the *Appendix II*.

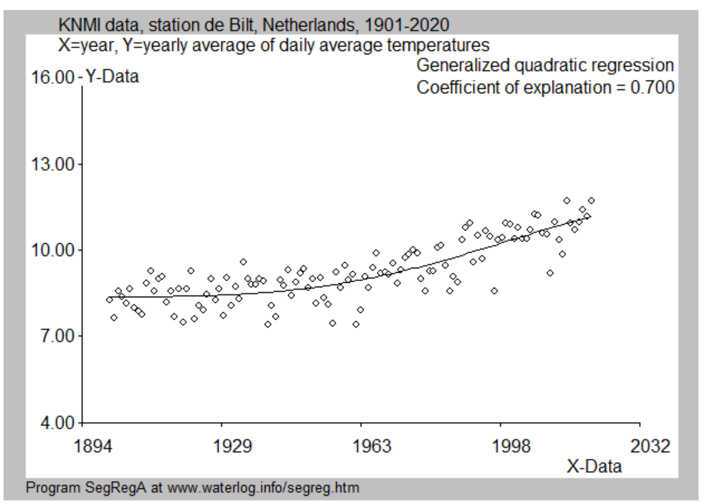


Figure 5. Temperature trend with fitted generalized quadratic curve with $R^2 = 0.700$
$Y = -6.87{*}10^{-13}\ W^{2*2.97} + 2.90{*}10^{-6}\ W^{2.97} + 8.38$
where $W = X - 1.894$

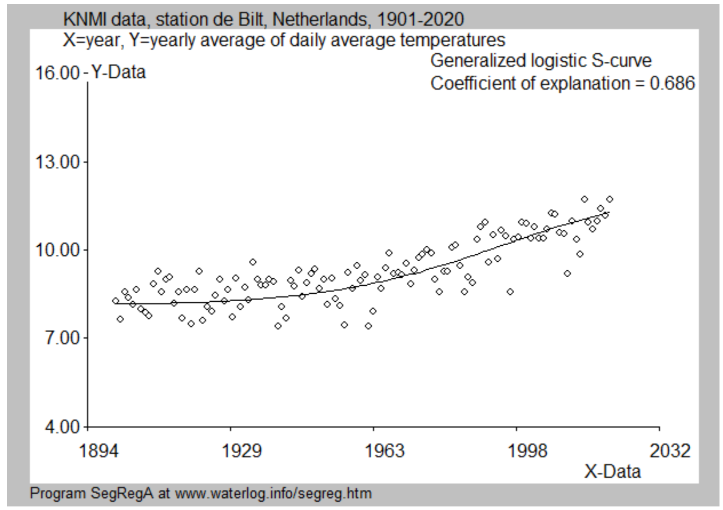


Figure 6. Temperature trend with the generalized logistic S-curve with $R^2 = 0.686$
$Y = 4.45\,/\,[1 + \exp(-8.93{*}10^{-6}\ W^{2.21} + 1.47)] + 7.35$
where $W = X - 1901$

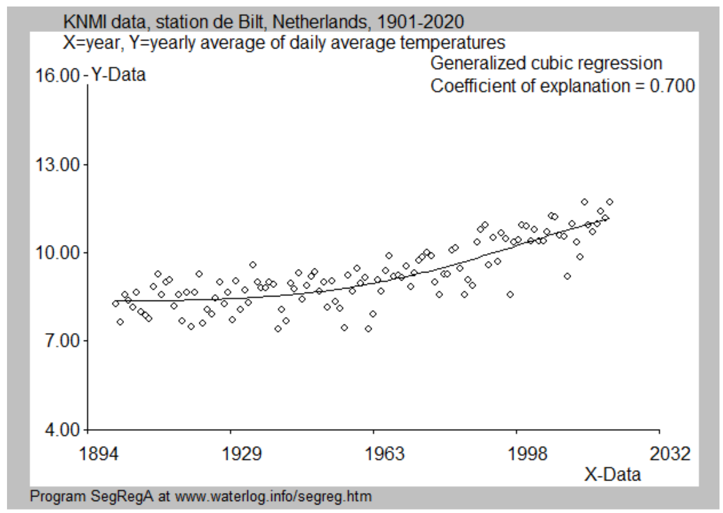


Figure 7. Temperature trend with fitted generalized cubic curve with $R^2 = 0.700$
$Y = -1.32{*}10^{-19}\ W^{3*2.82} - 2.23{*}10^{-12}\ W^{2*2.28}$
$+ 5.59{*}10^{-6}\ W^{2.82} + 8.30$ *where* $W = X - 1.894$

All the coefficients of explanation (R^2) are very close to 0.70 and do not provide a clear criterion for the best fit. To select the most appropriate curved function, it could be recommended to adopt the one with the least number of parameters, being the power curve [*Ref. 7*].
Also, all curves demonstrate an increasing trend owing to global warming, but the S-curve (*figure 6*) indicates that the increasing trend is slowing down at the end somewhat. If that is physically a logical feature (which, however, still has to be proved) then perhaps the S-curve deserves preference over the power curve. The decision is up to the user considering the environmental situation.

4. Analyzing the relation between yield of sugar cane and depth of water table, Australia

The results for the four curved regression types available in SegRegA are shown in respectively *figure 8, 9, 10 and 11*.

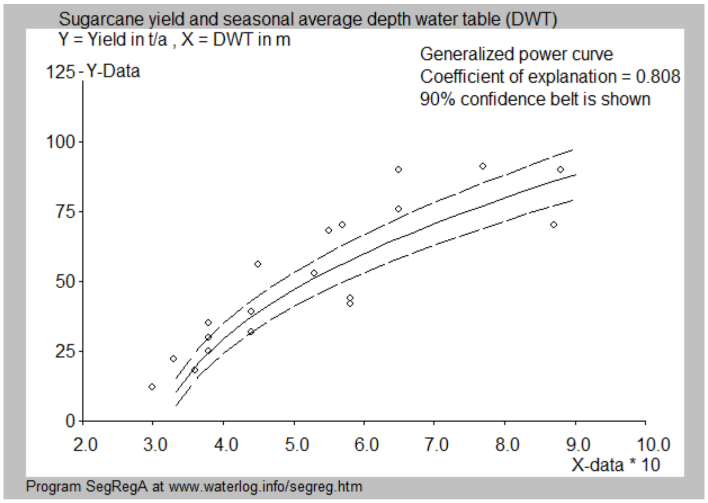


Figure 8. Power curve for the relation between yield and depth of water table:

$Y = 136 (X - 0.273)^{0.43} - 21.0$ *with* $R^2 = 0.808$

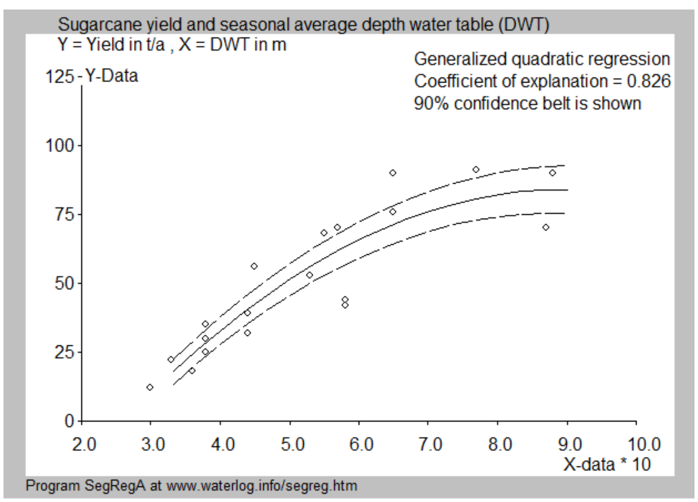


Figure 9. Generalized quadratic curve for the relation between yield and depth of water table:

$Y = - 0.021 W^2 + 249 W + 9.88$

with $W = X - 0.268$ *and* $R^2 = 0.826$.

In this case the option of the use of the transformation with the exponent E has been eliminated. Hence the quadratic function has not been generalized.

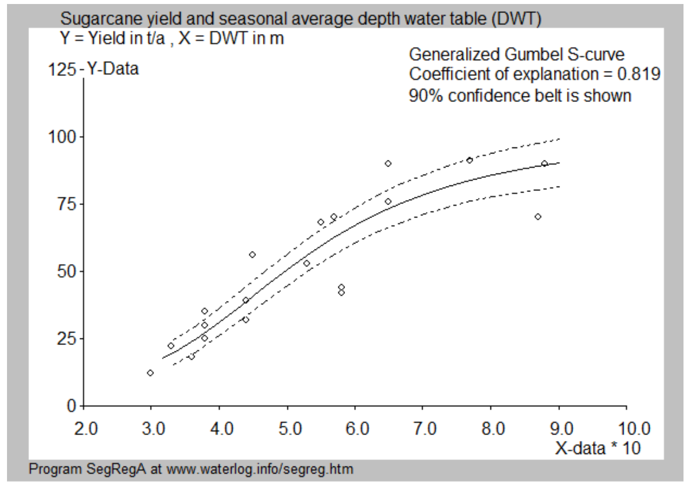


Figure 10. Generalized Gumbel S-curve for the relation between yield and depth of water table:

$Y = 84.7 \exp [- \exp (-5.19 X^{0.86} + 11.1)] + 11.4$

with $R^2 = 0.819$

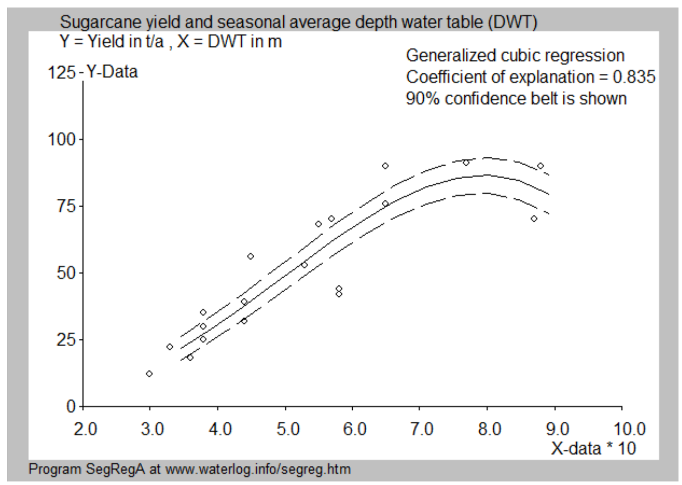


Figure 11. Generalized cubic curve for the relation between yield and depth of water table:

$Y = - 640 W^3 + 262 W^2 + 184 W + 16.3$

with $W = (X-0.268)^{1.86}$ and $R^2 = 0.835$

The quadratic (*figure 9*) shows that the initial rising trend is decreasing at deeper water tables and ultimately the slope will be zero. This corresponds to the agricultural experience that deeper water tables do not any more affect the yield negatively. Still more pronounced, the generalized cubic curve (*figure 11*) depicts a yield reduction at deeper water tables, which is perhaps not logical. The quadratic curve, theoretically will also yield a downward trend at water tables deeper than 1 m. Taking this into account, the S-curve (*figure 10*) seems to be the most recommendable final choice, even though its coefficient of explanation is less than that of the cubic function.

Note that the generalized cubic function *(figure 11*) is not exactly cubic any more. In Excel one needs the Solver add-in to accomplish the transformation, but the solver solution has its limitations [*Ref. 9*].

5. Analyzing the relation between potato yield and soil salinity, The Netherlands

The results for the four curved regression types available in SegRegA are shown in respectively *figure 12, 13, 14 and 15*.

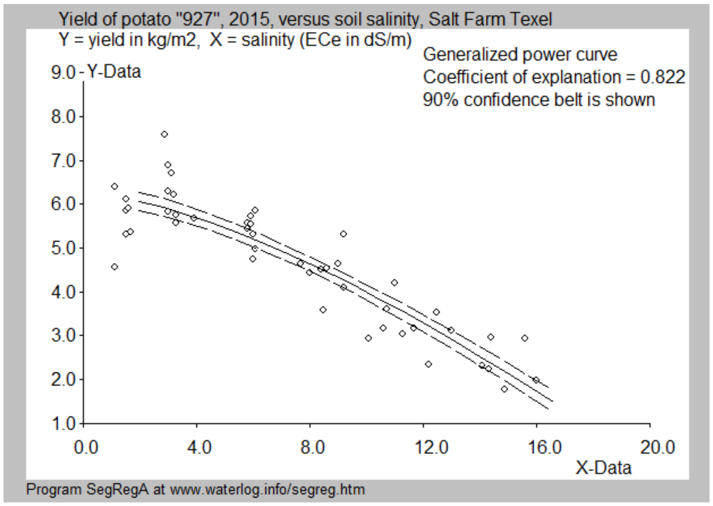


Figure 12. Power curve for the relation between yield and soil salinity:
$Y = -0.103\ (X-1.09)^{1.39} + 6.14$ with $R^2 = 0.822$

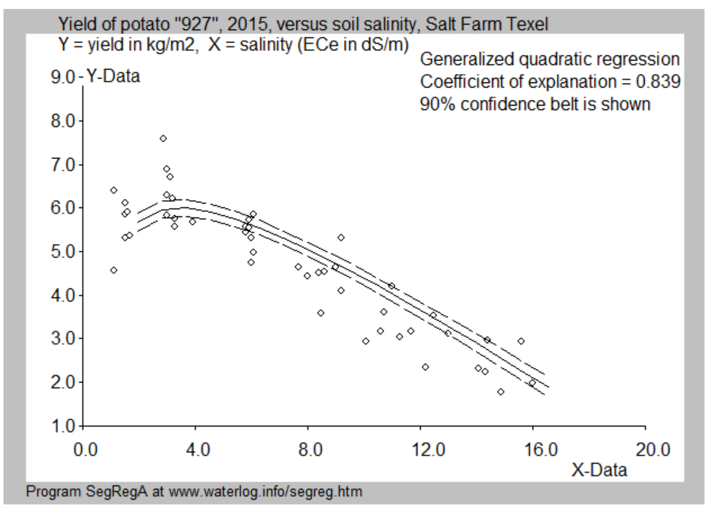


Figure 13. Generalized quadratic curve for the relation between yield and soil salinity:
$Y = -0.009\ X^2 - 0.0146\ X + 6.38$ with $R^2 = 0.814$

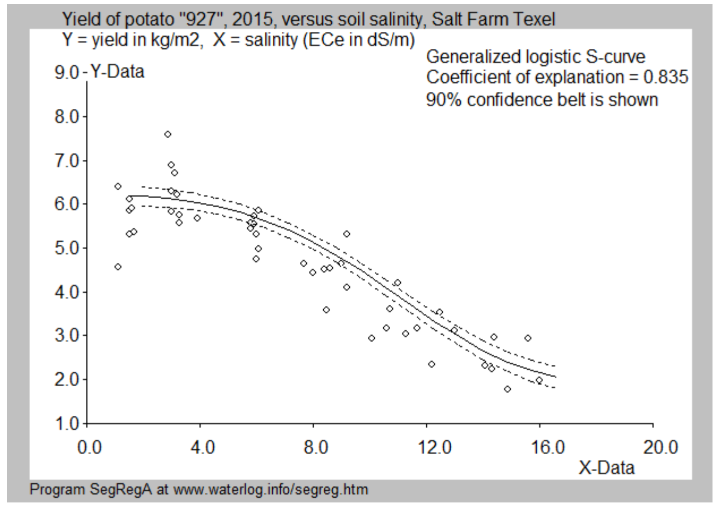


Figure 14. Generalized logistic S-curve for the relation between yield and soil salinity:
$Y = 7.00 / [1 + \exp(0.0174\ X^{1.98} - 1.07)] + 1.73$
with $R^2 = 0.835$

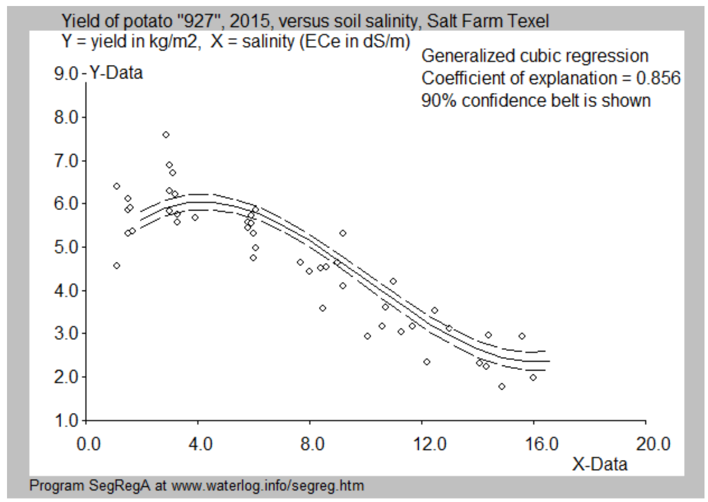


Figure 15. Generalized cubic curve for the relation between yield and soil salinity:
$Y = 0.0525\ W^3 - 0.768\ W^2 + 2.65\ W + 3.49$
with $W = X^{0.73}$ and $R^2 = 0.864$

The generalized cubic curve (*figure 15*) has the highest goodness of fit ($R^2 = 0.864$) and also shows the logical plant physiological phenomenon that initially the yield increases with increasing soil salinity because plants do need some salts anyway. Only at the higher salinity levels the yield starts declining due to the excess salinity.

It would be recommendable to select the generalized cubic regression as the definitive choice. This is contrary to the previous two cases (*Sections 3 and 4*) where the Power curve and the S-curve deserved preference.

Compared to *figure 3* in Section 1 (Introductions) showing a segmented regression, The cubic regression deserves preference as one can expect the relation between crop yield and soil salinity to be fluent without abrupt change.

Note that here, like in Section 3, the generalized logistic S-curve is used, whereas in Section 4 the generalized Gumbel S-curve is found on the basis of its higher R^2 value.

6. Summary of results

All cases were checked with Analysis of Variance (ANOVA) to confirm that the curve functions give a statistically significant improvement over a linear (straight line) function. In SegRegA this check occurs automatically.

6.1 Analyzing the average temperature trend (Section 3)

All the coefficients of explanation (R^2) are very close to 0.70 and do not provide a clear criterion for the best fit. To select the most appropriate curved function, it could be recommended to adopt the one with the least number of parameters, being the Power curve [*Ref. 7*].

All curves demonstrate an increasing trend owing to global warming, but the S-curve (*figure 6*) indicates that the increasing trend is slowing down at the end somewhat. If that is physically a logical feature (which, however, still has to be proved) then perhaps the S-curve deserves preference over the power curve. The decision rests with the user.

6.2 Analyzing the relation between yield of sugar cane and depth of water table (Section 4)

The quadratic (*figure 7*) shows that the initial rising trend is decreasing at deeper water tables and ultimately the slope will be zero. This corresponds to the agricultural experience that deeper water tables do not any more affect the yield negatively. Still more pronounced, the generalized cubic curve (*figure 11*) depicts a yield reduction at deeper water tables, which is perhaps not logical. The quadratic curve, theoretically will also yield a downward trend at water tables deeper than 1m. Taking this into account, the S-curve (*figure 10*) seems to be the most recommendable final choice, even though its coefficient of explanation is less than that of the cubic function.

6.3 Analyzing the relation between potato yield and soil salinity (Section 5)

The generalized cubic curve (*figure 15*) has the highest goodness of fit (R^2 = 0.864) and also shows the logical plant physiological phenomenon that initially the yield increases with increasing soil salinity because plants do need some salts. Only at the higher salinity levels the yield start declining due to the excess salinity. It would therefore be a good choice.

7. Conclusions

It has been found that all four different curve functions have given good results. Once the power curve has been preferred, once the S-curve and once the cubic function.

In the case of the temperature analysis (*Section 3*), it has been argued that the choice between the Power curve and the S-curve must depend on the knowledge of climatological conditions.

The final choice is mainly determined by the value of the explanation coefficient R^2, but there may exist experiences that the final selection is not made on the grounds of its maximum value. In this respect, the statistical confidence belt of R^2 can play a decisive role [Ref. 10]. The confidence interval of R^2 is discussed in *Appendix III.*. In addition, as the Analysis of Variance ANOVA could play a role..

Examples of the testing of the differences between quadratic and cubic regression based on ANOVA are to be found in reference 11.

It may be stipulated that the curved regression is not always a recommendable solution as it does not always give a significant improvement over a straightforward linear regression [Ref. 9]. As an example, abuse of the S-curve, is discussed in *Appendix IV*.

Finally, in the various examples given, it has been stipulated that regressions are purely mathematical tools, and that logical considerations or environmental situations need also be considered before deciding on the final choice of the curved regression type.

All in all, no strict rules can be given for the determination of the final choice

References

[1] G.J. van Oldenborgh and G. Komen, 2001. *Does the warm weather of the latest years in The Netherlands correlate with the global warming*? (*in Dutch*). In: Meteorologica, Year code 10 Nr. 3, September 2001. On line: https://congressus-nvbm.s3-eu-west-1.amazonaws.com/files/339e10da367040458314258d1f96931e.pdf

[2] Article preprint published in Research gate on Sept. 2019. On line: The potato variety "927" tested at the Salt Farm Texel,The Netherlands, proved to be highly salt tolerant or: https://www.waterlog.info/pdf/Potato 927

[3] SegRegA, free software for segmented and curved regressions. On line: https://www.waterlog.info/segreg.htm

[4a] S-curves for fitting measured crop growth to measures growth factors. On line: https://www.waterlog.info/sigmoid.htm

[4b] Free software for the determination of positive and inverted S-curves for the response function of influential treatments or conditions with examples of crop yield versus soil salinity and depth of the water table, 2021. Technical report published in Researchgate. On line: Free calculator for the determination of positive and inverted S-curves for the response function of influential treatments or conditions with examples of crop yield versus soil salinity and depth of the water table or: https://www.waterlog.info/pdf/S-curves.pdf

[5] Fitting the versatile linearized, composite, and generalized logistic probability distribution to a data set. Conference paper published in Researchgate on April 2021. On line: FITTING THE VERSATILE LINEARIZED, COMPOSITE, AND GENERALIZED LOGISTIC PROBABILITY DISTRIBUTION TO A DATA SET or: https://www.waterlog.info/pdf/logistic.pdf

[6] Software for generalized and composite probability distributions, 2019. In: International Journal of Mathematical and Computational Methods, 4, 1-9. On line: https://www.iaras.org/iaras/home/caijmcm/software-for-generalized-and-composite-probability-distributions or: https://www.waterlog.info/pdf/MathJournal.pdf

[7] Comparing the regressions of Y-X data by means of the amplified power function using Solver in Excel and SegRegA, 2021. Conference paper published in Research gate On line: Comparing the regressions of Y-X data by means of the amplified power function using Solver in Excel and SegRegA or: https://www.waterlog.info/pdf/Power function.pdf

[8] Rudd, A.V. and C.W Chardon 1977. *The effects of drainage on cane yields as measured by water table height in the Machnade Mill area*. In: Proceedings of the 44th Conference of the Queensland Society of Sugar Cane Technology, Australia

[9] Questionable mirrored S-curves used in literature on crop yield relations with soil salinity to determine salt tolerance of crops, 2021. Technical report published in Researchgate. On line: Questionable mirrored S-curves used in literature on crop yield relations with soil salinity to determine salt tolerance of crops or: https://www.waterlog.info/pdf/Strange Scurves.pdf

[10] Free software for finding the confidence belt of R^2 values. On line: https://www.waterlog.info/r-squared.htm

[11] Testing the statistical significance of the improvement of cubic regression compared to quadratic regression using analysis of variance (ANOVA), 2020. Article preprint published in Research gate in 2020. On line: Testing the statistical significance of the improvement of cubic regression compared to quadratic regression using analysis of variance (ANOVA) or: https://www.waterlog.info/pdf/Anova cubic.pdf

APPENDIX I. Screen shots of the input menu input SegRegA

Figure I below gives a screen shot of the input menu.

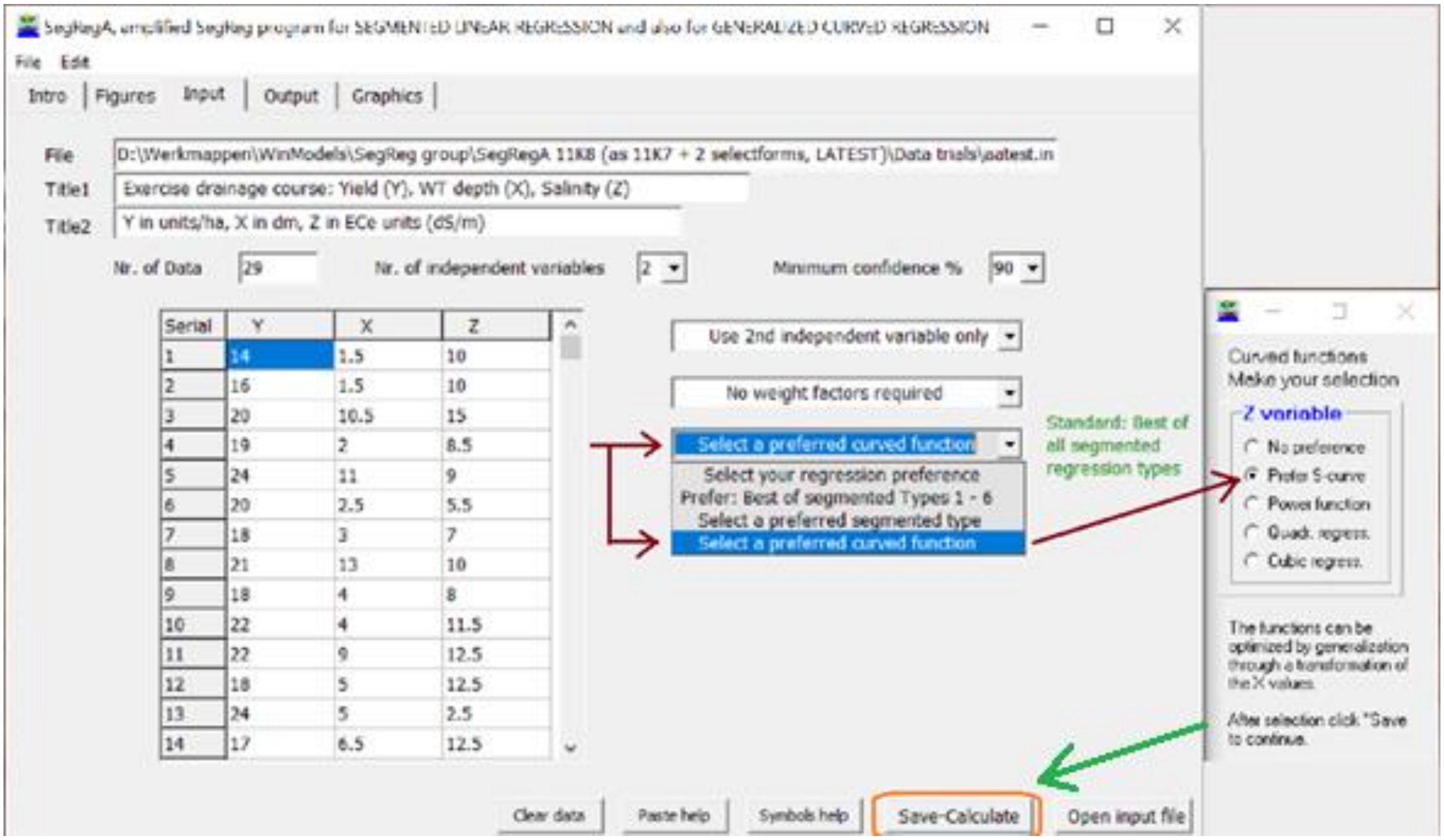


Figure I. Selection options in the input menu (user interface) of SegRegA. Here, for example, the S-curve option is selected (arrows). This option produces either the logistic, or the Gumbel, or the Gumbel mirrored S-curve, whichever gives the highest goodness of fit. After the selection the Save-Calculate button is to be used (green arrow).

APPENDIX II. Output screen of SegRegA showing parameters of curve functions

Figure II below gives a screen shot of the output menu

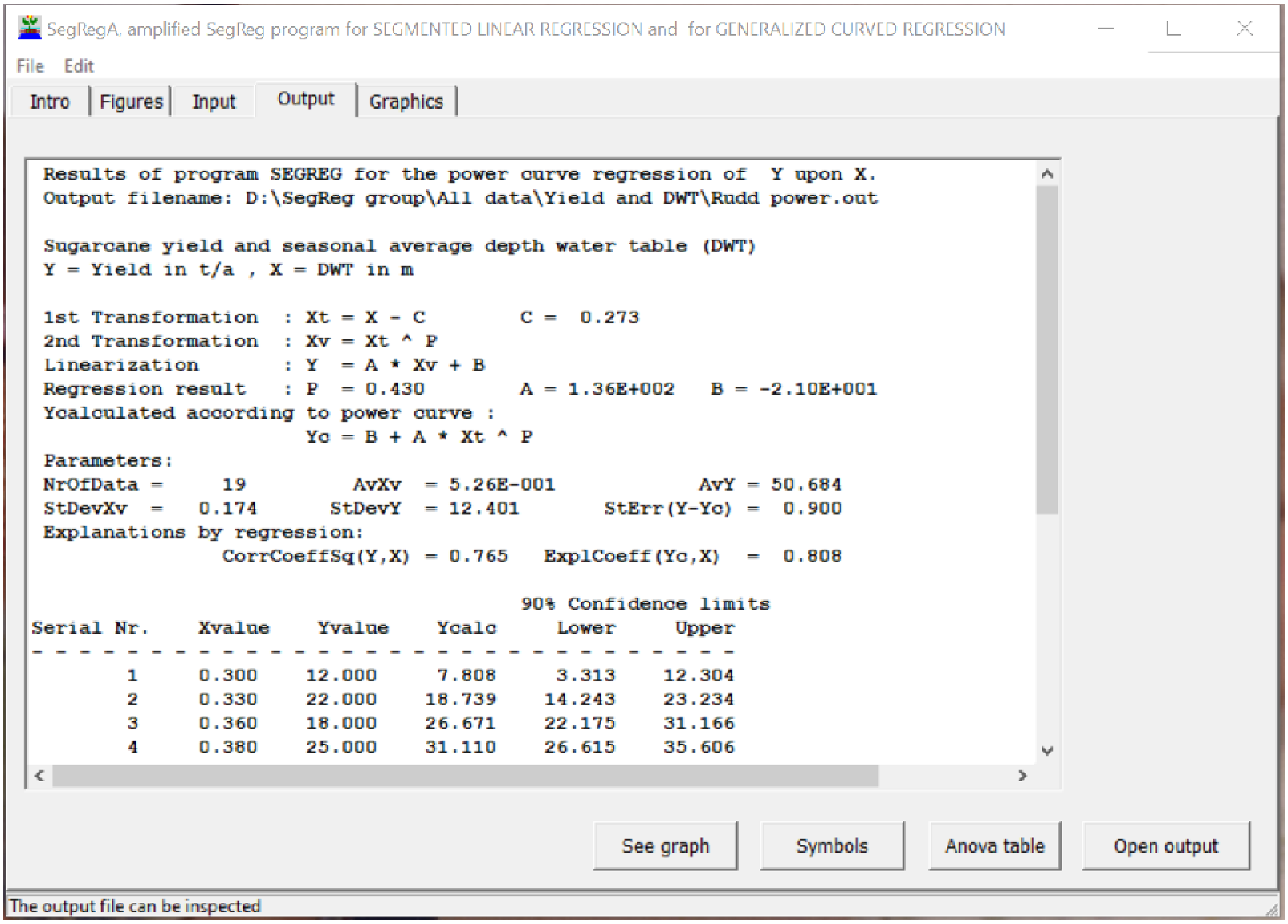
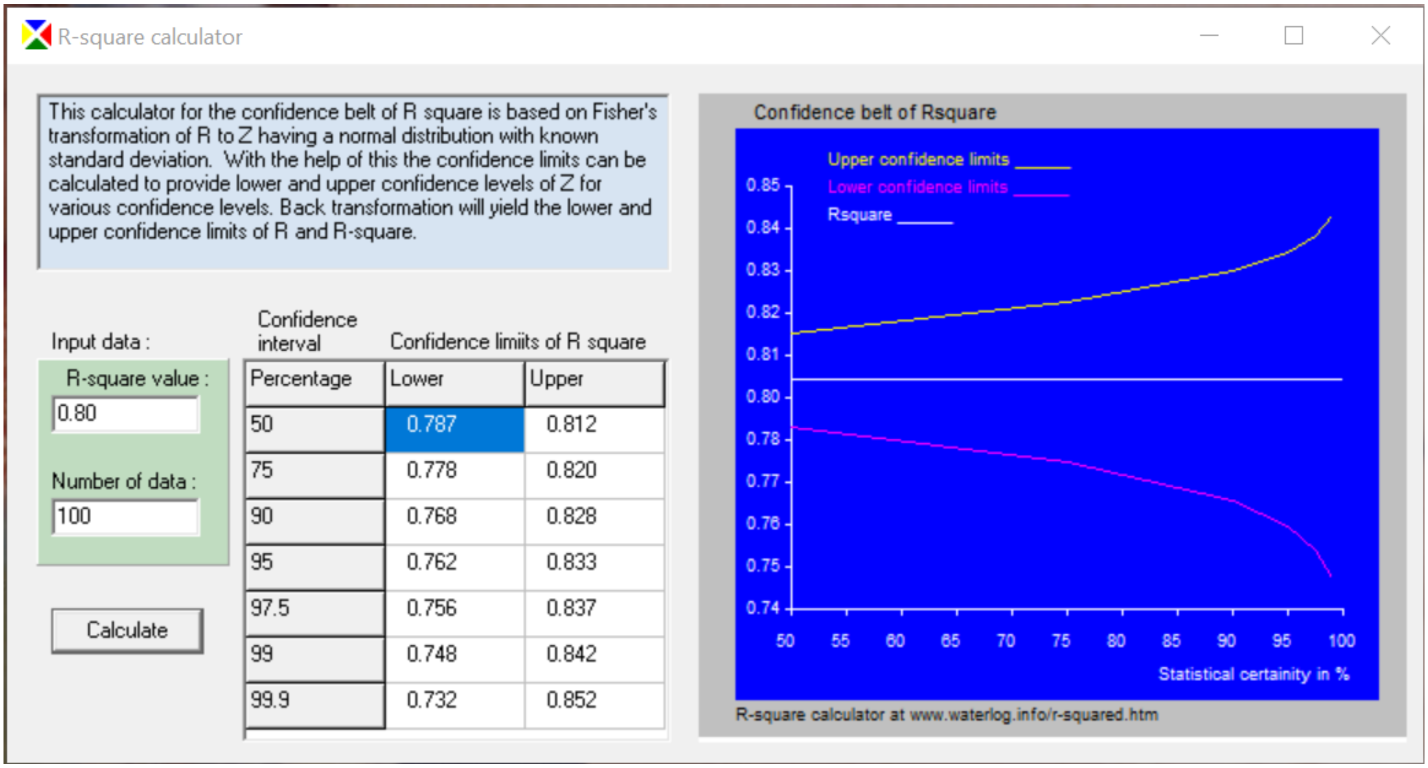


Figure II, Screen shot of the output menu of SegRegA showing the transformations, the final equation and parameter values. This example concerns the sugar cane yield versus depth of the water table (DWT) dealt with in *Section 4*.

APPENDIX III. Free software for the determination of the confidence belt of the index R^2 for the goodness of fit of curved functions to the

The R-square calculator to find the confidence belt of R^2 is freely available from https://www.waterlog.info/r-squared.htm
The mathematics are also described in it.
A screen shot of it is given hereunder.



When the R^2 value is 0.80 and the number of data sets is 100, the 90% confidence interval of R^2 ranges between 0.768 and 0.828. When two values of R^2 are, for example, found as 0.79 and 0.81 they are statistically not significantly different. However, when they are 0.75 and 0.83, they are different, except when a higher security limit (for instance 99%) is adopted, then the difference is no longer significant unless the two values are <0.748 and > 0.842 respectively.

APPENDIX IV. Questionable use of S-curves and an ANOVA (analysis of variance) table

In this appendix questionable S-curves have been reported in:

> "Questionable mirrored S-curves used in literature on crop yield relations with soil salinity to determine salt tolerance of crops". On line:
> https://www.researchgate.net/publication/349074586_Questionable_mirrored_S-curves_used_in_literature_on_crop_yield_relations_with_soil_salinity_to_determine_salt_tolerance_of_crops
> or:
> https://www.waterlog.info/pdf/Strange Surves.pdf

The data used stem from different publications that will be referred to in continuation.

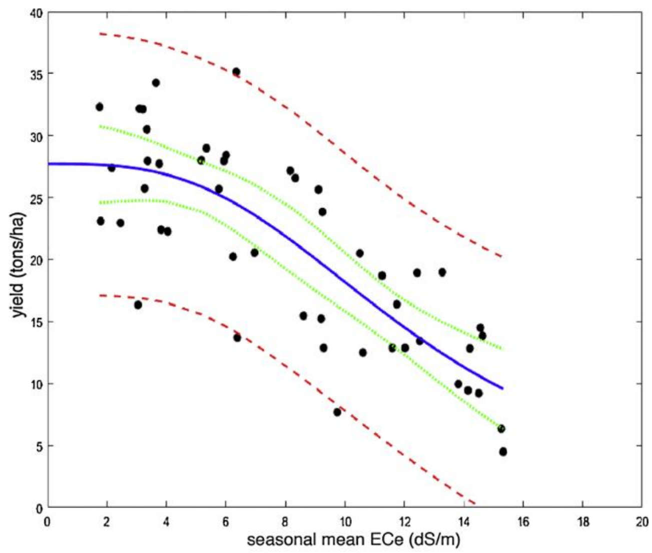
The figure pasted in the table below stems from:

> *G. van Straten et al., 2019.* An improved methodology to evaluate crop salt tolerance from field trials. *In: Agricultural Water Management, Volume 1, March 2019, Pages 375-387. On line:* *https://www.sciencedirect.com/science/article/pii/S0378377418310370*

The right hand figure is copied from the Van Straten publication mentioned above and concerns the yield of potato Achilles tested in 2014 versus soil salinity (ECe in dS/m).

The figure is based on the van Genuchten-Hoffman S-curve model *) , also known as the van Genuchten- Gupta *) S-curve model

*) The respective publications are cited hereunder



Analysis of variance (ANOVA, in SegRegA) of the van Genuchten (VG) model to test statistically the improvement of this model with respect to a simple linear regression in the Achilles 2014 case.

The F-test provides a significance level of only 48%, meaning that there is 52% chance that the VG model has arisen by coincidence and that the model does significantly improve the straightforward linear regression,

The use of the VG model is not justified and linear regression is preferable.

Sum of squares of deviations	Degrees of freedom	Variance	Fisher's F-test	Probability, significance (%)
Total 44.200	47	0.940		
Explained by lin. regress. 25.300	1	25.300	F(1,46)= 61.577	99.9 %
Remaining Unexplained 18.900	46	0.411		
Extra explained by vGG model 0.533	2	0.271	F(2,44)= 0.666	48.1 %
Remaining Unexplained 18.375	44	0.408		

The publications on which the van Genuchten (VG) model is based are the following:

M. Th. Van Genuchten and S.K. Gupta, 1993. *Reassessment of the Crop Tolerance Response Function*. Journal of the Indian Society of Soil Science, Vol. 41, No. 4, pp 730-737 (1993). On line: https://www.ars.usda.gov/ARSUserFiles/20360500/pdf_pubs/P1295.pdf?origin=publication_detail

M. Th. Van Genuchten and G.J Hoffman, 1984. *Analysis of crop salt tolerance data*. On line: https://www.researchgate.net/publication/238185339_Analysis_of_crop_salt_tolerance_data

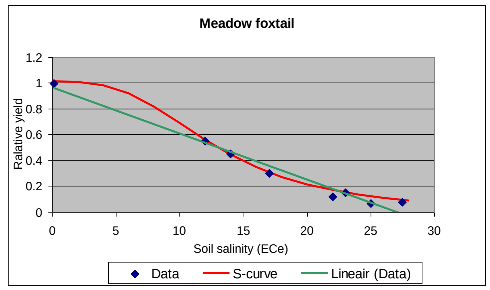
The next two figures are from the same sources.

Van Genuchten and Gupta:

Relative yield of meadow foxtail versus soil salinity (ECe in dS/m). The deviation of the S-curve from the straight line, therefore, has no experimental justification because there are no data in this part.

Beyond X = 12 dS/m, the differences between the S-curve and the straight line are negligibly small. The use of the S-curve in this case is not justified.

The correspondence between the S-curve and the observations is amazingly high. This may give rise to the question: have the data been manipulated?

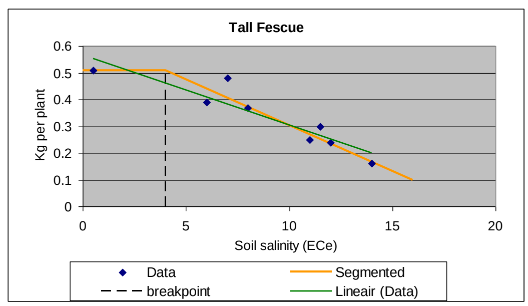


The red line shows the fitted S-curve, while the green line represents a simple linear regression. In the range between X = 1 dS/m to X = 12 dS/m there are no data. #)

Van Genuchten and Hoffman

Relation between yield of tall fescue in kg/plant (Y) and soil salinity (ECe in dS/m, X). The comments made above are equally relevant here.

The determination of the breakpoint (at X between 4 and 5) as a tolerance index is questionable as the straight line (in green color) seems to be statistically the most appropriate solution here.



#) Plants normally grow well in soils with an ECe value less than 6 dS/m, while beyond that value the yields decline. It is strange that in this figure the data (except one) are all in the uninteresting range of low production.