Regression by Integration applied to Ångström-Prescott-type relations

Heinrich Morf

Senior Member ISES, Buechraiweg 47, 5452 Oberrohrdorf (Switzerland)

Abstract

We present a novel approach for the determination of the relationship between two random variables, which we call *Regression by Integration*. The resulting curve is a least absolute error estimate. Compared to other regression methods, it has the advantage that, instead of a sample of simultaneously taken pairs of the two random variables, only a separate sample of each of the random variables is required. We demonstrate the practicability of the method on Ångström-Prescott-type relations and compare the results with those obtained by least square error fits.

Keywords: Ångström-Prescott relation, curve fit, Regression by Integration, random variable

1. Introduction

Representing the relationship between two random variables – *X* and *Y* by a curve is established engineering practice. Solar energy engineering routinely employs many such relations.

A well-established procedure is the least square error regression, where a curve *y(x)* is fitted to a sample of pairs of values *(x,y)* that minimizes the sum of the error squares in either the direction of *x* or *y*. (See for example Kreiszig (2006), pp. 860-863.) While the objective of the fit is clear, the result is somewhat ambiguous as it depends on the choice of the variable (*x* or *y*) whose least square error was minimized.

We present an alternative method that treats the two random variables equally, which we call *Regression by Integration*. It has the advantage that – instead of a sample of pairs of values *(x,y)* – only separate samples of *x* and *y* need to be known. It is a least absolute error estimate; the probability for an outcome of *(x,y)* to be on either side of the regression curve is the same. As a least absolute error estimate, it is less susceptible to outliers than a least square error fit.

This paper is organized as follows. First, the basic mathematical model leading to *Regression by Integration* is described. Then, we introduce to the properties of the model that are important when applying the method. A step-by-step application procedure follows. Its practicability is demonstrated on examples of Ångström-Prescott-type relations.

2. Introduction into Regression by Integration

We give a succinct introduction into *Regression by Integration*. We present the basic model behind the method. Then, we introduce to the properties of the model that are important when applying the method.

2.1. Regression by Integration – the idea

The method of *Regression by Integration* builds on the expression of the relationship between two strictly dependent random variables – *Y(X)* as an ordinary differential equation (See for example Brandt, 1968, pp. 27-29.):

$$f_2(y) \cdot dy = f_1(x) \cdot dx \qquad \text{eq. (1)}$$

with the initial condition

$$y(x = x_{min}) = y_{min} \qquad \text{eq. (2)}$$

$f_1(x)$ and $f_2(y)$ are the probability density functions (pdfs) of *X* and *Y*.

doi:10.18086/eurosun2018.09.10 Available at http://proceedings.ises.org

The objective is to find the function *y(x)*.

We assume that *y(x)* will be a non-decreasing function. (Situations resulting in a decreasing function can be transformed into the non-decreasing case by a variable transformation.) Therefore, eq. (1) can be solved by equalizing the distribution functions of $X - F_1(x)$ and $Y - F_2(y)$, and state

$$F_1(x) = F_2(y) = F^*, \qquad \text{eq. (3)}$$

whereby F^* may be any value within its range of validity $0 \leq F^* \leq 1$.

$$F_1(x) = \int_{x=x_{min}}^{x} f_1(x) \cdot dx = P(X \leq x) \qquad \text{eq. (4)}$$

and

$$F_2(y) = \int_{y=y_{min}}^{y} f_2(y) \cdot dy = P(Y \leq y) \qquad \text{eq. (5)}$$

The solution is then

$$y(x)|F^* = (x(F^*), y(F^*)) \qquad \text{eq. (6)}$$

eq. (6) states that *x* and *y* at F^* are a pair of values that obey *y(x)*.

We will also employ eq. (1) to find an estimate for *y(x)* when there is a bivariate distribution of the two random variables – *(X,Y)* where the data points *(x,y)* form rather a data cloud than a curve in the *x,y* plane. Our estimate is then the relationship between the two random variables under the assumption of strict interdependence.

2.2. Key features of Regression by Integration

We point to some properties of the method and its results. In first place we do so, because their understanding is essential for its successful application. In second place we would like to substantiate the advantages of *Regression by Integration* in comparison to other methods. For the sake of clarity, we restrict our exposition to the minimum necessary for the application of the method. Justifying details and theoretical proofs may be found in Morf (2018).

- The fit is a least absolute error estimate for the random variable *(X,Y)*. The probability for an outcome *(x,y)* to be on either side of the regression curve is the same.
- Instead of a sample of *(x,y)*, only separate samples of *x* and *y* need to be known. This increases the number for prospective sample choices, because samples of *X* and *Y* taken over different time periods and even at different locations may be considered.
- When the standardized distributions of *X* and *Y* are the same, the fit leads to a linear relation of the form $y = a + b \cdot x$.
- Often, the result deviates little from a least square error fit. However, this is not necessarily so. For example, applying *Regression by Integration* to two random variables *X* with $0 < x \leq 1$ and *Y* with $0 < y \leq 1$ expanding a uniform bivariate distribution, the result would be $y = x$, whereas for a linear least square error regression it would be $y = 0.5$.
- As a least absolute error estimate the result is less susceptible to outliers than a least square error estimate.

3. Application of the method

We present a step-by-step application procedure for *Regression by Integration*. The practicability of the procedure is demonstrated on examples of Ångström-Prescott-type regressions over various time periods.

3.1. A procedure for the generation of y(x)

We devised the following step-by-step procedure for the generation of pairs of values *(x,y)* of *y(x)*:

- Take samples of *X* and *Y* in the real world.

- Generate discrete probability functions of the samples – $f_1(x_i)$, respectively $f_2(y_j)$ – by sorting the outcomes into classes of appropriate width delimited by $x_1, x_2, \ldots x_i, \ldots x_m$, respectively $y_1, y_2, \ldots y_j, \ldots y_n$.
- Generate discrete distribution functions of X and Y:

  $F_1(x_i) = P(X \le x_i) = \sum_{x_i \le x} f(x_i)$, and

  $F_2(y_j) = P(Y \le y_j) = \sum_{y_j \le y} f(y_j)$

- Estimate the continuous inverse functions $x(F_1)$ and $y(F_2)$ using polynomial least square error regression, or any other method that suits the data sets at hand.
- Set $F_1(x)=F_2(y)=F^*$. Then $(x(F^*),y(F^*))=(x,y)$.

From many values (x,y) generated in the last step above, one may then generate a continuous fit for $y(x)$, if required. (Least square error, least absolute error, cubic spline, or any other fit suiting a specific problem at hand.)

3.2. Applications on Ångström-Prescott-type relations

We present applications of *Regression by Integration* on Ångström-Prescott-type relations over various time periods. They relate the clearness index to the relative sunshine duration over the given time period. We refer to the nomenclature for definitions of these two random variables.

In the presented graphs results of *Regression by Integration* are marked with triangles tagged with the corresponding common value of the marginal distributions – F^*. For comparison, the results of linear and quadratic least square error fits are also shown.

3.2.1. Monthly fits for daily values

Fig. 1 depicts the results of *Regression by Integration* on the relation between the daily clearness index – K and daily relative sunshine duration – S, for Payerne, Switzerland and Perth, Australia in January, April, July, and October, based on ten-year datasets.

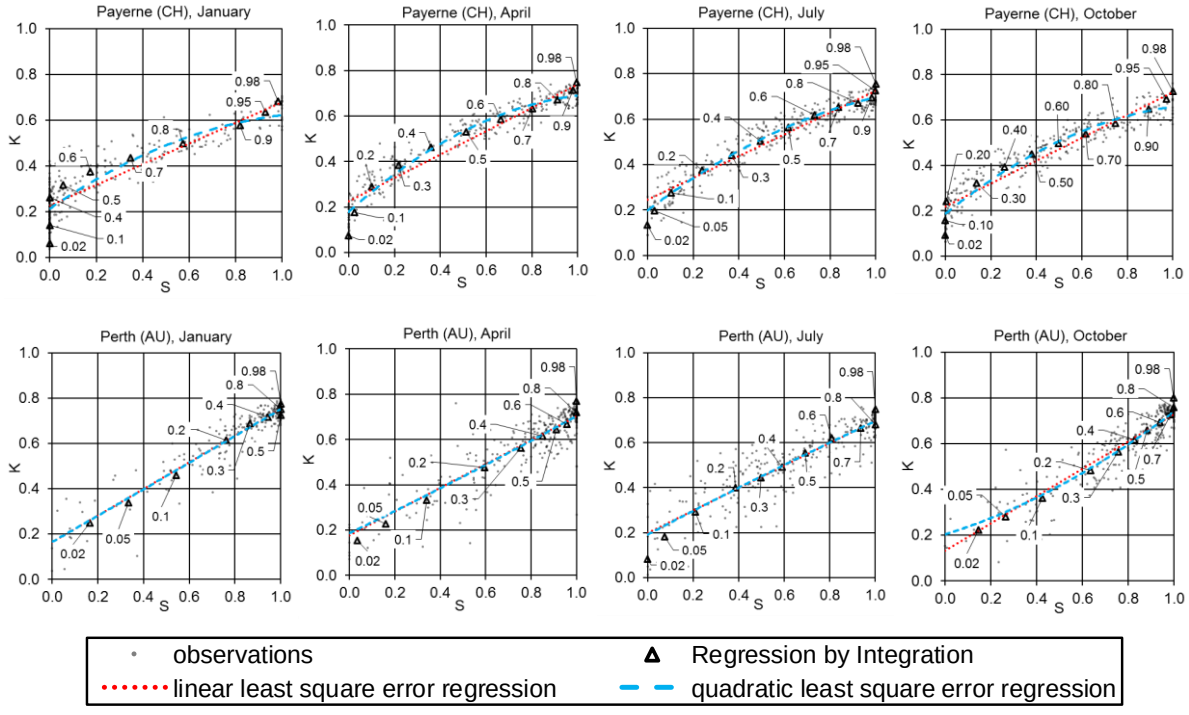

Fig. 1: ***Regression by Integration*** **for the daily clearness index –** ***K*** **as a function of daily relative sunshine duration –** ***S*****, for Payerne and Perth in January, April, July, and October, based on ten-year data-sets.**

The points for *Regression by Integration* were determined following the procedure outlined in Section 3.1. The fits rise from $(K=0,S=0)$, bend over to $S=1$ where they continue to rise to $K=1$. Due to the inaccuracy of the results we refrain from presenting fitted data points for the lower and upper two percent of F^*. Nevertheless, the point for $F^*=0$ must be located at $(K=0, S=0)$, because K and S cannot be negative. Also on physical grounds, we place $F^*=1$ at $(K=1, S=1)$.

Fitting a curve through the results of *Regression by Integration*, will hardly ever result in a linear fit for daily values. Scrutiny at *S=0* reveals that the attenuation of solar irradiation by clouds is variable. Having a closer look at *S=1*, one concludes that also clear sky irradiation is of varying intensity.

3.2.2. Monthly fits for monthly means

Fig. 2 depicts the result of *Regression by Integration* on the relation between monthly means of daily clearness index – $\bar{K}$ and monthly means of daily relative sunshine duration – $\bar{S}$, for Payerne, Switzerland and Perth, Australia in January, April, July, and October, based on ten-year datasets.

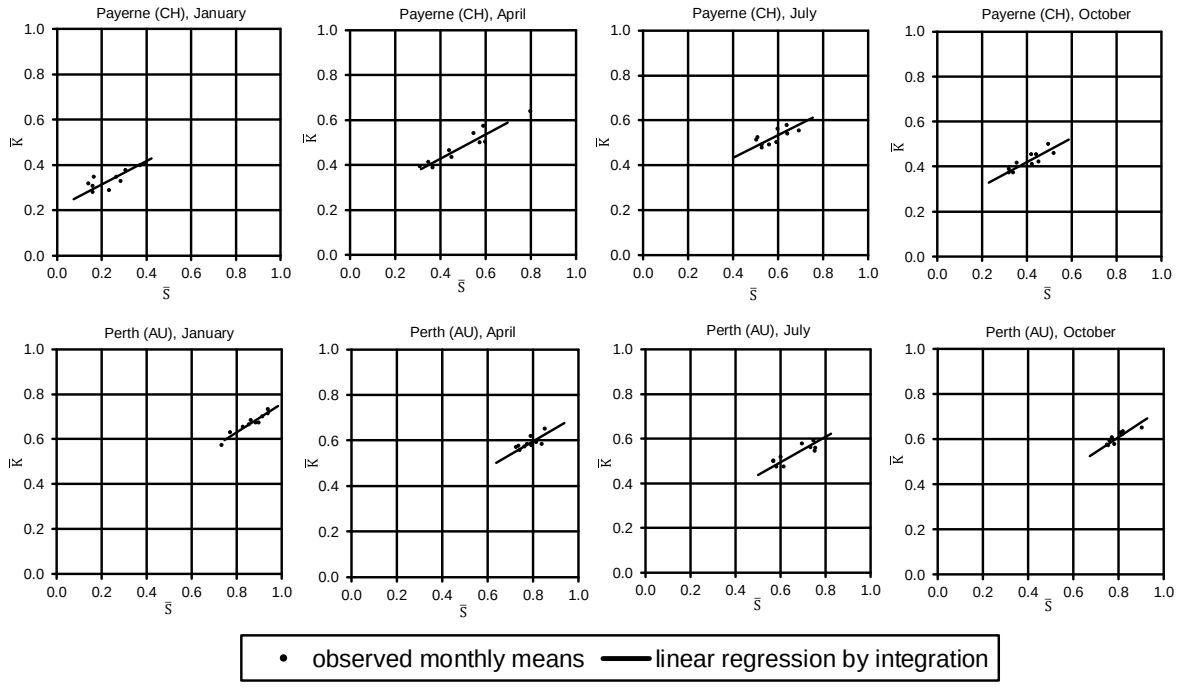

Fig. 2: ***Regression by Integration*** **for the monthly means of the daily clearness index –** $\bar{K}$ **as a function of the monthly means of daily relative sunshine duration –** $\bar{S}$**, for Payerne and Perth in January, April, July, and October, based on ten-year data-sets. The line is the** ***Regression by Integration*** **calculated under the assumption that the relation be linear. It expands over the range of** $\mu \pm 3\sigma$ **of the distributions of** $\bar{K}$ **and** $\bar{S}$**. 99.7% of the outcomes of a normal variable would be within this range**

Being the mean of roughly 30 daily values, monthly means present a probing ground for linear fits of *Regression by Integration*. Based on the Central Limit Theorem (See for example Kreiszig, 1998, pp. 200-202.) one concludes that – under ideal conditions – means of the daily clearness index and means of daily relative sunshine duration are asymptotically normal. Consequently, the standardized distributions of $\bar{K}$ and $\bar{S}$ will be equal, as required for a linear outcome of *Regression by Integration*.

One obtains for the linear fit of *Regression by Integration*

$$\bar{K} = a + b \cdot \bar{S} \qquad \text{eq. (7)}$$

with intercept

$$a = \mu_2 - \frac{\sigma_2}{\sigma_1} \cdot \mu_1 \qquad \text{eq. (8)}$$

and slope

$$b = \frac{\sigma_2}{\sigma_1} \qquad \text{eq. (9)}$$

μ_1 and σ_1 are mean and standard deviation of *daily* relative sunshine duration – *S*. μ_2 and σ_2 are mean and standard deviation of *daily* clearness index – *K*.

Working with monthly means is not an ideal situation to obtain linear fits. Contiguous daily values of K and S are not completely independent. The number of values – n to form the means is limited to about 30. In statistical data analysis, a value of $n=30$ is considered to be sufficiently large to obtain reliable results for the mean, whereas for the variance a value of $n=100$ is deemed to be necessary (Kreiszig, 1998, pp. 201-202). Nevertheless, a special Kolmogorov-Smirnov test (Lilliefors, 1967) applied to test whether the standardized distributions of $\bar{K}$ and $\bar{S}$ were normal seldom failed, and the standard two sample Kolmogorov-Smirnov test (Kreiszig, 1998, pp. 234-237) applied to test the equality of the standardized distributions of $\bar{K}$ and $\bar{S}$ always passed. The test results are listed in Table 1.

Table 1. Kolmogorov-Smirnov tests on monthly means of the daily clearness index – $\bar{K}$ and monthly means of daily relative sunshine duration – $\bar{S}$, for Payerne and Perth in January, April, July, and October, based on ten-year data-sets. The standardized distributions were tested for normality (KS-1) and equality of the corresponding standardized distributions of $\bar{K}$ and $\bar{S}$ (KS-2). A significance level of 0.01 was used for the Kolmogorov-Smirnov tests. Sample size: 10 for the monthly tests, and 120 for the yearly tests. Also mean and variance of the data sets are given. The regression parameters for the linear *Regression by Integration* were calculated from these values.

Period	$\bar{S}$			$\bar{K}$			$\bar{K}$&$\bar{S}$	Regression	
	Mean	Variance	KS-1	Mean	Variance	KS-1	KS-2	*a*	*b*
PAYERNE									
Jan	2.480E-01	3.410E-03	fail	3.385E-01	8.390E-04	pass	pass	0.2122	0.5091
Apr	5.047E-01	4.116E-03	pass	4.851E-01	1.203E-03	fail	pass	0.2122	0.5407
Jul	5.802E-01	3.361E-03	pass	5.236E-01	8.484E-04	pass	pass	0.2311	0.5024
Oct	4.088E-01	3.523E-04	pass	4.242E-01	1.013E-03	pass	pass	0.2050	0.5362
Year	4.305E-01	2.568E-02	pass	4.414E-01	7.928E-03	pass	pass	0.2022	0.5556
PERTH									
Jan	8.664E-01	1.590E-03	pass	6.721E-01	6.323E-04	pass	pass	0.1258	0.6305
Apr	7.887E-01	2.507E-03	pass	5.899E-01	8.533E-04	pass	pass	0.1297	0.5835
Jul	6.625E-01	2.961E-03	fail	5.297E-01	9.323E-04	pass	pass	0.1580	0.5611
Oct	7.999E-01	1.806E-03	pass	6.084E-01	7.710E-04	pass	pass	0.0858	0.6533
Year	7.741E-01	1.144E-02	pass	5.973E-01	4.979E-04	pass	pass	0.0868	0.6596

3.2.3. Yearly fits for monthly means

Fig. 3 depicts the result of *Regression by Integration* on the relation between the monthly average daily clearness index – $\bar{K}$ and monthly average daily relative sunshine duration – $\bar{S}$, for Payerne, Switzerland and Perth, Australia over the entire year based on ten year datasets.

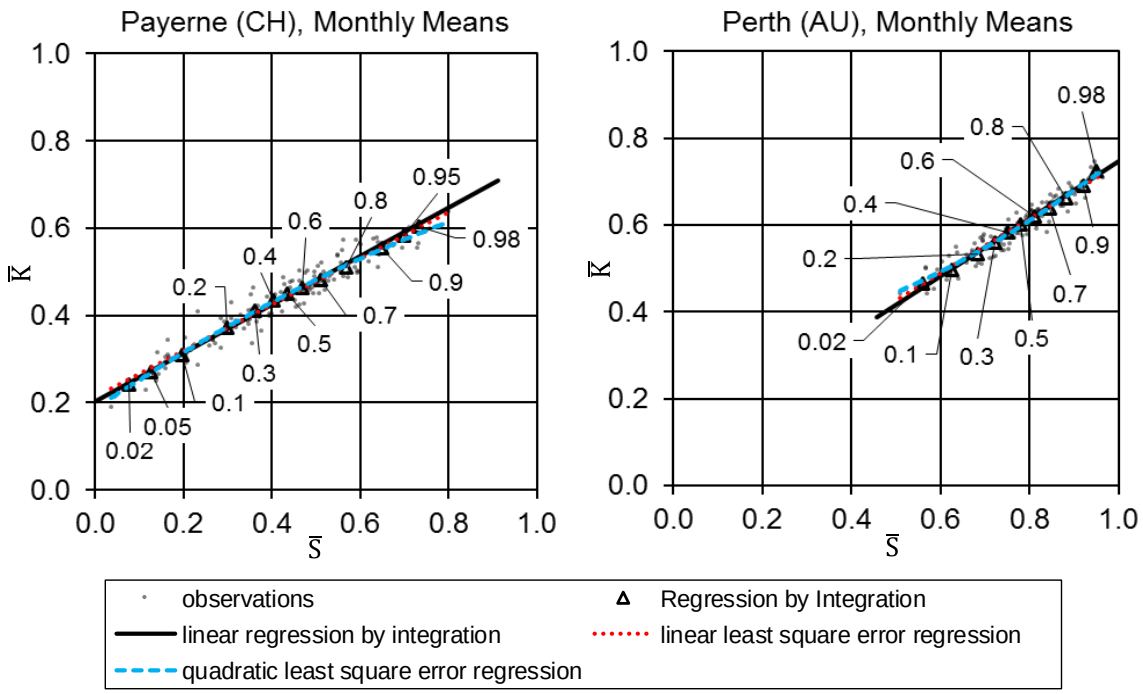

Fig. 3: *Regression by Integration* for the monthly means of the daily clearness index – $\bar{K}$ as a function of the monthly means of daily relative sunshine duration – $\bar{S}$, for Payerne and Perth taken over the entire year, based on ten-year data sets. The full line is the *Regression by Integration* calculated under the assumption that the relation be linear. It expands over the range of μ ± 3σ of the distributions of $\bar{K}$ and $\bar{S}$. 99.7% of the outcomes of a normal variable would be within this range.

The data points were determined by *Regression by Integration* as outlined in Section 3.1. The linear fits of *Regression by Integration* were calculated under the assumption that the relation be linear. The linear least square error fit, the quadratic least square error fit, and the linear fit by *Regression by Integration* can hardly be distinguished on the graphs.

Kolmogorov-Smirnov tests to verify whether the standardized distributions of $\bar{K}$ and $\bar{S}$ were normal and equal always passed. Results are listed in Table 1. As may be seen, monthly fits (Fig. 2) and fits over the entire year (Fig. 3) deviate little.

4. Multivariate distributions

Expanding the method of *Regression by Integration* to multivariate distributions is a natural next step in pursuit of the perfect fit. For a multivariate distribution one expands eq. (3) to

$$F_1(1) = F_2(2) = F_3(3) \ldots = F_n(n) = F^* \qquad \text{eq. (10)}$$

(For flexibility, we substituted the alphabetic mnemonics for the dimension in eq. (3) by their corresponding numbers.)

Fig. 4 depicts trivariate distributions of daily clearness index – *K*, daily relative sunshine duration – *S*, and daily mean cloud cover – *C*. One observes the curves of the bivariate distributions *K(S)*, *K(C)*, and *C(S)*. The intercept of the projection of points with equal F^* value of the three two-dimensinal curves (marked with bullets) leads to the corresponding point of the three-dimensional curve (marked with a triangle and tagged with the corresponding value of F^*).

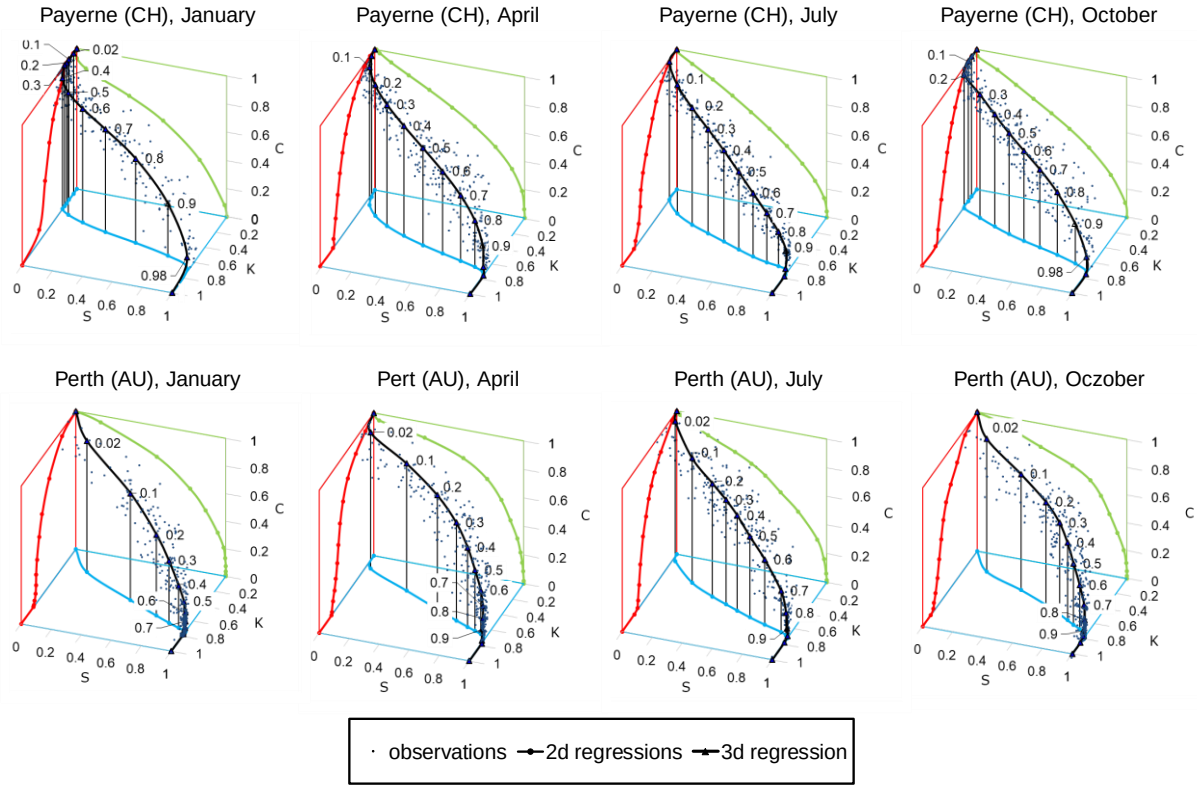

Fig. 4: ***Regression by Integration*** **applied to the trivariate distribution of daily clearness index –** ***K*****, daily relative sunshine duration –** ***S*****, and daily mean cloud cover –** ***C*** **for Payerne and Perth in January, April, July, and October, based on ten-year data-sets.**

5. Concluding remarks and outlook

In this paper we present a novel method for the determination of the relationship between two random variables, which we call *Regression by Integration*. We demonstrate its practicability on Ångström-Prescott-type relations and compare the results with those obtained by least square error fits.

The method treats the two random variables equally. As to be expected for the dependence of two strictly dependent stochastic variables, the result will always be a monotonic function. Compared to other regression methods, it has the advantage that – instead of a sample of simultaneously taken pairs of the two random variables – only a separate sample of each of the random variables is required. This increases the number of prospective sample choices, because samples of the two random variables may be taken over different time periods, and even at different locations.

Morf (2018) presents justifying details and proofs for many of the properties of *Regression by Integration* based on copulas. Among others, he demonstrates that

- *Regression by Integration* is a least absolute error estimate: the probability for an outcome to be on either side of the regression curve is the same. As a least absolute error estimate, it is less susceptible to outliers than a least square error fit.
- *Regression by Integraton* leads to the curve $y(x)$ of strict interdependence beween the two stochastic variables X and Y; Spearman's rho is eaqual to one.

The method is universally applicable. Treating the two random variables equally, and being a least absolute error estimate, it will be the best choice for many applications.

6. Data

Publicly available data of the following locations have been used:

- Payerne, Switzerland, 46.80° N / 6.93° E, for the years 2001-2010.
- Perth Airport, Australia, 31.93° S / 115.98° E, for the years 1998-2007.

From both stations records of daily clearness index, daily relative sunshine duration, and 3-hourly synoptic cloud cover observations were used. Daily mean cloud cover was calculated as the mean of the synoptic cloud cover observations from sunrise to sunset, using linear interpolation. The rare days with unreliable or missing data were removed.

7. Acknowledgements

The author wishes to thank the Western Australian Climate Service Centre for the supply of a ten year data set of Perth Airport, containing daily clearness index, daily relative sunshine duration, and 3-hourly synoptic total cloud cover observations.

The author wishes to thank the Swiss Federal Office of Meteorology and Climatology – Meteo Swiss for the supply of a ten year data set of Payerne, containing daily clearness index, daily relative sunshine duration, and 3-hourly synoptic total cloud cover observations.

8. References

Brandt, S., 1968. Statistische Methoden der Datenanalyse. Bibliographisches Institut, Mannheim.

Kreiszig, E., 1998. Statistische Methoden und ihre Anwendungen. 7th Edition. Vandenhoek und Ruprecht, Göttingen.

Kreiszig, E., 2006. Advanced Engineering Mathematics. Ninth Edition. Wiley and Sons, Hoboken, NJ.

Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62, 399-402.

Morf, H., 2018. Regression by Integration demonstrated on Ångström-Prescott-type of relations. Renewable Energy 127, 713-723. https://doi.org/10.1016/j.renene.2018.05.004

9. Nomenclature

Capital Letters

C	daily mean cloud cover The mean of cloud cover over the period between sunrise and sunset.
F	probability distribution (function)
K	daily clearness index Ratio between daily solar irradiation [W s m^{-1}] on a horizontal surface and daily extraterrestrial irradiation on a horizontal surface [W s m^{-1}].
P	probability
S	daily relative sunshine duration Ratio between daily sunshine duration [s] and the time between sunrise and sunset [s].
X	independent random variable
Y	dependent random variable

Small Letters

a	intercept of a linear regression
b	slope of a linear regression
f	probability function of a discrete random variable
f	probability density function of a continuous random variable
n	number of days in a calendar month, sample size
x	independent variable
y	dependent variable

Greek Letters

μ	mean, expectation
ρ	Spearman's rho
σ	standard deviation
σ^2	variance

Special Characters

$\|$	Condition
$\bar{}$	monthly mean
*	equal value of two marginal probability distribution (function)

Indices

1	of the independent random variable (mostly of X)
2	of the dependent random variable (mostly of Y)
i	generic index
j	generic index
min	minimum
max	maximum