Sampling Theory for Continuous Variables in Space and Time

Hartmut Hebbel

hartmut.hebbel@unibwh.de

Detlef Steuer

steuer@unibwh.de

University of the Federal Armed Forces Hamburg

14.08.2003

Problem

1. Duration curve

2. Space-time sampling designs

3. Directions of research

Literature

Problem

This researchs stems from an application in hydrology.

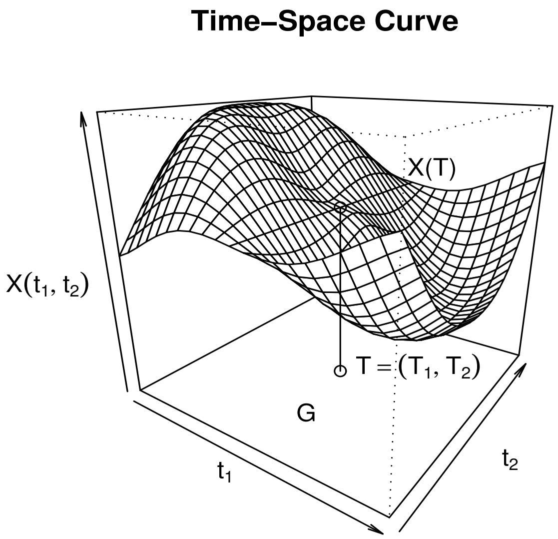
How to estimate annual loads

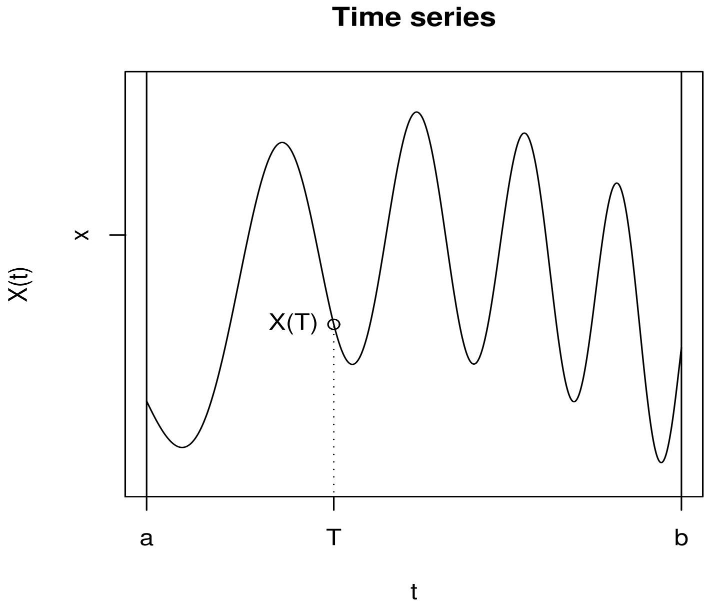
$$\int_G C(t)Q(t)\,dt$$

of an ingredient with concentration $C(t)$[mg/l] and river flow $Q(t)$[m^3/s] on a given cross-section of a river for a year, if only some measurements are available?

Formal description

Measurement $X : G \to \mathbb{R}$, G area or interval $G = [a, b]$





$T \in G$ "purely random", i.e. T uniformly distributed in G with density

$$f_T(t) = \frac{1}{|G|} \quad \text{for} \quad t \in G \quad \text{and} \quad 0 \quad \text{else;} \quad |G| \quad \text{volume of} \quad G.$$

$\Rightarrow \quad X(T)$ is random variate.

Question

What is the distribution of the "random measurement" $X(T)$?

Known: For a finite population G, $X(T)$ is distributed according to relative frequencies.

$$P\big(X(T) = x_i\big) = \frac{n_i}{n}$$

Is there an analogy for a continuous (infinite) population?

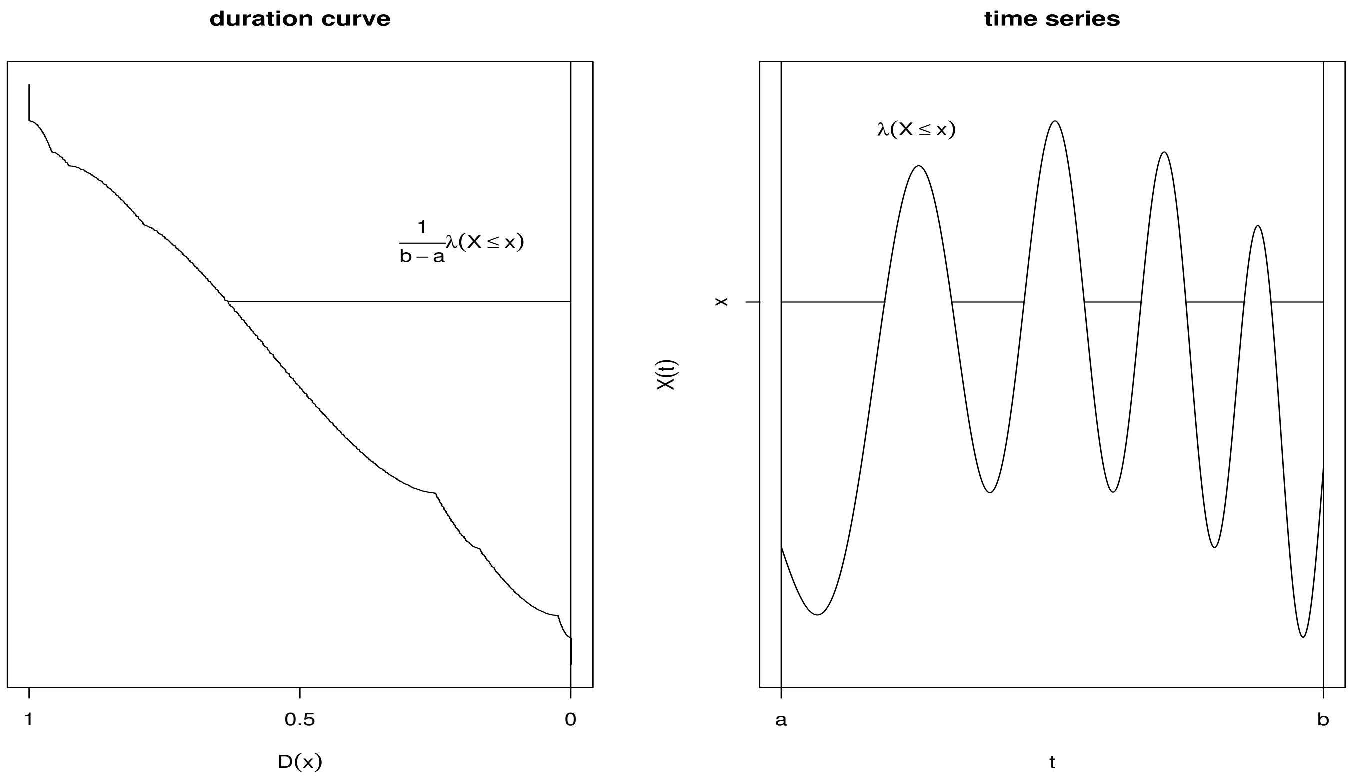
1. Duration curve

Idea: replace relative frequencies with portions of area resp. interval G.

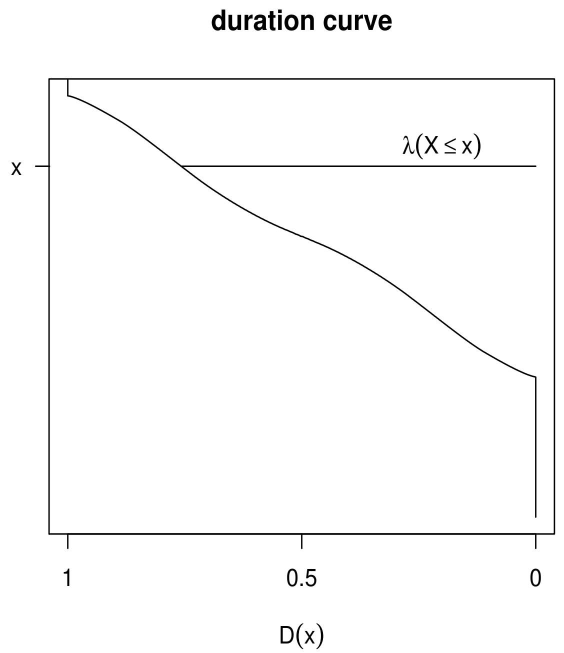
Definition The duration curve of $X : G \to \mathbb{R}$ is defined as

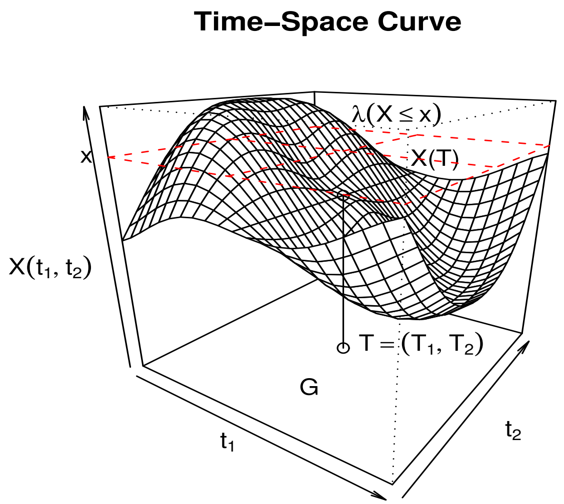
$$D(x) = \frac{1}{|G|}\lambda(X \leq x)\,, \quad x \in \mathbb{R}$$

case 1: (time-) interval $G = [a, b]$



case 2: area G; may be irregular shaped





Definition The joint duration curve of $X_1, \ldots, X_k : G \to \mathbb{R}$ is defined as

$$D(x_1, \ldots, x_k) = \frac{1}{|G|}\lambda(X_1 \leq x_1 \ldots, X_k \leq x_k)\,, \quad x_1, \ldots, x_k \in \mathbb{R}\,.$$

portion of area or time where jointly $x_1, \ldots, x_k$ are not exceeded.

Properties of the duration curve

- monotonically non decreasing and

$$D(x) = 0 \quad \text{for} \quad x \leq \min_{t \in G} X(t)$$

$$D(x) = 1 \quad \text{for} \quad x \geq \max_{t \in G} X(t)$$

- In general D is not differentiable in extreme points x_E of X. Therefore no analogy for a density exisits.
- The Lesbesgue-integral over X in G can be written as Riemann-integral in terms of λ :

$$\int_G X(t)\, dt = \int_0^\infty \Big(\underbrace{|G| - \lambda(X \leq x)}_{\lambda(X > x)} \Big) dx \; - \underbrace{\int_{-\infty}^0 \lambda(X \leq x)\, dx}_{\text{if } X \text{ also} \leq 0}$$

Remark

Empirical duration curve

$$D_n(x) = \frac{1}{|G|} \sum_{i:X(t_i)\leq x} |G_i|$$

for (fixed) observation points $t_i \in G$ with subareas resp. subintervals G_i, $i = 1, \ldots, n$, which build a partition of G,

$$g_i = \frac{|G_i|}{|G|} \quad \Rightarrow \quad \sum_{i=1}^{n} g_i = 1$$

D_n formally resembles the empirical distribution function if $g_i = \frac{1}{n}$, but has a different interpretation: here portions of area or length, there portions of numbers of values.

This formal resemblence has lead to a common error in literature: the empirical duration curve is approximated with a known distribution function, interpreting X as random variate, which it is not.

Some hydrologist came near to reinvent the Lesbesgue-integral !

2. Space-time (Time) sampling designs

Definition A sampling design, where an n-tupel $(t_1, \dots, t_n)$ of observation points is chosen randomly from G is called time sampling design, written as $(T_1, \dots, T_n)$, with $T_i, i = 1, \dots, n$ being random variates.

The following is central:

Theorem For a randomly chosen observation point T in G the joint distribution function of (random) measurements $X_1(T), \dots, X_k(T)$ is given by the joint duration curve, i.e.

$$P\big(X_1(T) \le x_1, \dots, X_k(T) \le x_k\big) = D(x_1, \dots, x_k)\,, \quad x_1, \dots, x_k \in \mathbb{R}$$

Proof $\Updownarrow$

$$P\big(T \in \{t \in G \,|\, X_1(t) \le x_1, \dots, X_k(t) \le x_k\}\big) = \underset{\uparrow \text{ uniform distribution of T}}{\frac{1}{|G|}}\lambda(X_1 \le x_1, \dots, X_k \le x_k)$$

According to the definiton of the expected value

$$\mathrm{E}\big(h(T)\big) = \int_{\mathbb{R}} h(t) f_T(t)\, dt = \frac{1}{|G|}\int_G h(t)\, dt\,, \quad h \quad \text{measurable}$$

moments exist, without restriction $k = 2$ with X, Y:

Corollary

$$\mathrm{E}\big(X(T)\big) = \frac{1}{|G|}\int_G X(t)\, dt =: \mu_X\,, \quad \text{“mean over interval”, “mean over area”}$$

Examples: mean temperature for a month, mean rainfall for a day in an area

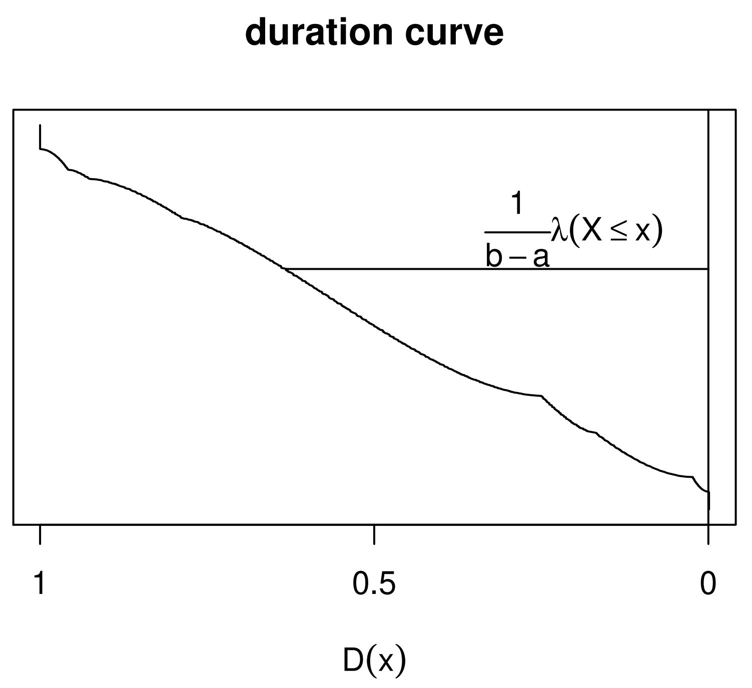
$$\begin{aligned}\mathrm{Var}\big(X(T)\big) &= \mathrm{E}\Big((X(T) - \mu_x)^2\Big) \\ &= \frac{1}{|G|}\int_G \big(X(t) - \mu_X\big)^2\, dt =: \sigma_X^2\,, \quad \text{“variance over interval/area”}\end{aligned}$$

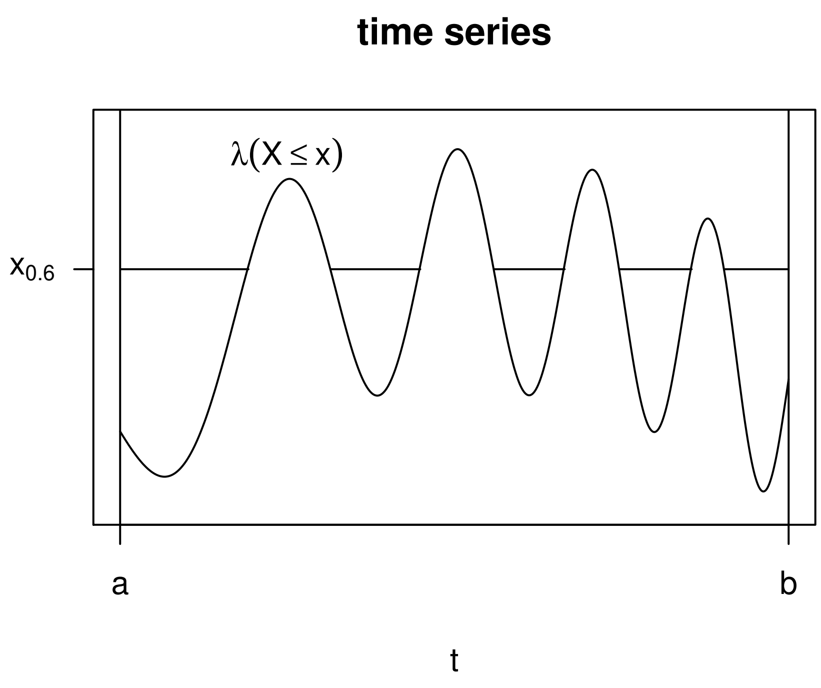
$$\mathrm{Cov}\big(X(T), Y(T)\big) = \frac{1}{|G|}\int_G \big(X(t) - \mu_X\big)\big(Y(t) - \mu_Y\big)\, dt =: \sigma_{XY}\,, \ \text{“Int.cov.”}$$

Duration quantiles

x_p where $D(x_p) = p$ "quantile of duration", "quantile of area"

Example: $p = 0.6$. In 60% of time the value $x_{0.6}$ of X is not exceeded in G and in 40% it is.





2.1 Random time sampling design (resp. X, Y)

There exist special sampling designs to estimate μ_X, σ_x^2, σ_{XY} etc. (analogous to classical theory of finite populations).

Definition Random time sampling design, iff

$$(T_1, \ldots, T_n) \quad \text{iid uniformly over } G$$

Theorem Then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X(T_i) \quad \text{is BLUE for } \mu_X$$

with

$$\mathrm{Var}(\overline{X}) = \frac{1}{n}\sigma_X^2\,, \quad \mathrm{Cov}(\overline{X}, \overline{Y}) = \frac{1}{n}\sigma_{XY}$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X(T_i) - \overline{X}\right)\left(Y(T_i) - \overline{Y}\right) \quad \text{is BQUE for } \sigma_{XY}$$

with

$$\mathrm{Var}(S_{XY}) = \ldots \qquad \mathrm{Cov}(S_{XY}, \overline{X}) = \ldots$$

etc.

2.2 Stratified time sampling design

Definition Given a partition G_i with lengths $|G_i|$, $i = 1, \ldots, m$ of G. You get a stratified time sampling design, if in each of the strata G_i (with$\overline{X}_i$, $\mu_X^{(i)}$, $\sigma_X^{(i)}$, $\sigma_{XY}^{(i)}$) a random time sampling design

$$(T_{i1}, \ldots, T_{in_i}) \quad \text{with } n_i \text{ observation times, } \sum_{i=1}^{m} n_i = n,\ g_i = \frac{|G_i|}{|G|}$$

is applied.

Theorem Then

$$\widehat{\mu}_X = \sum_{i=1}^{m} \overline{X}_i\, g_i \quad \text{is BLUE for } \mu_X$$

with

$$\mathrm{Var}(\widehat{\mu}_X) = \sum_{i=1}^{m} \sigma_X^{(i)2} \frac{g_i^2}{n_i}, \quad \mathrm{Cov}(\widehat{\mu}_X, \widehat{\mu}_Y) = \sum_{i=1}^{m} \sigma_{XY}^{(i)} \frac{g_i^2}{n_i}$$

etc. for higher (internal, external) Moments.

2.3 Systematic time sampling design

Definition Given a partition G_i $i = 1, \ldots, m$ of G. All G_i with equal lengths $|G_i| = \Delta = \frac{1}{n}|G|$. If an observation time $T_1 = T$ is randomly chosen in G_1 and $T_i = (i-1) \cdot \Delta$, this is called a systematic time sampling design.

$$T_i = T + (i-1)\Delta\,, \quad i = 1, \ldots, n\,, \quad T_1 = T \quad \text{uniformly distributed on } G_1$$

Theorem Then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X(T_i) \quad \text{is unbiased for } \mu_X$$

and

$$\mathrm{Var}(\overline{X}) = \frac{1}{n}\sigma^2_{X,\mathrm{int}} + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=1}^{n-i} \underbrace{\gamma^{(i)}(j\Delta)}_{\text{covariance of } X \text{ in } G_i \text{ and } G_{i+j}}$$

similar to stratified time sampling design with $n_i = 1$, $g_i = \frac{1}{n}$.

3. Future Research

All of the classical finite population sampling theory can be transfered into the continuous case of sampling in space and time.

The theory is potentially useful for algorithms for numerical integration.

A theory for sampling for space-time continuous stochastical-processes resp. weak stationary processes can be developed.

Literature

Hebbel, H. (2000). Theorie der Frachtberechnung und -standardisierung in der Wassergütewirtschaft – Entwurf von Zeitstichprobenplänen –. In: Discussion Papers in Statistics and Quantitative Economics Nr. 91 (eds H. Hauptmann, W. Krumbholz). University of the Federal Armed Forces Hamburg. Hamburg.

Hebbel, H., Steuer, D. (2001). Sampling for Continuous (Space-time) Processes. In: Mathematical Statistics with Applications in Biometry, Festschrift in Honour of Prof. Dr. Siegfried Schach (eds J. Kunert and G. Trenkler), 277-291. Eul. Lohmar, Köln.