IMPROVED AIR VENTILATION RATE ESTIMATION BASED ON A STATISTICAL MODEL

M.Brabec[1], K.Jílek[2]

[1] National Institute of Public Health, Department of Biostatistics, Praha, Czech Republic

[2] National Radiation Protection Institute, Praha, Czech Republic

Correspondence to: mbrabec@szu.cz

Abstract

We will outline a new approach to air ventilation rate estimation from CO measurement data that is based on a state-space dynamic statistical model, allowing for quick and efficient estimation. Underlying computations are based on Kalman filtering, whose practical software implementation is easy. Key property is the model's flexibility, so that it can handle various artificial regimens of CO level manipulation. The model is semi-parametric in nature and hence it can handle time-varying ventilation rate in an efficient way. This is a major advantage, compared to some of the methods which are currently in practical use. After introducing the statistical model formally, we will demonstrate its performance on real data from routine measurements. Further, we will describe, how our approach can be utilized in a more complicated situation of great practical relevance, when time-varying air ventilation rate and radon entry rate are to be estimated simultaneously from contemporary radon and CO measurements.

1. Introduction

Tracer-gas technique is a powerful method for assessment of radon entry rate, in various indoor situations. Its great practical impetus stems from the fact that simultaneous measurements of radon and artificially introduced tracer gas (e.g. CO) permit to compute both radon entry rate and ventilation rate. This is true under certain conditions which can be satisfied by appropriate choice of experimental setup, Jílek (2003), Jílek, Brabec (2004). In our applications, ventilation rate might or might not be of interest per se. Even if it is not of interest, it acts as a nuisance parameter, when trying to estimate radon entry rate, however. Unless the ventilation is known or tamed by severe external assumptions (which are hardly satisfied in practice), it disturbs the radon entry rate estimation. Therefore, it is not feasible to estimate the Rn entry rate from radon measurements alone. Simultaneous tracer and radon measurements can deal with this obstacle nicely, when taken under carefully selected experimental design. In this paper, we will deal with the design based on measurement in a closed but naturally ventilating room with artificial CO input of known constant rate (maintained by a precise pump). It is clear that when building a formal model, the particular choice of the tracer gas is not important. We have been working with other experimental setups as well (e.g. zero tracer input after certain tracer level has been reached). Modification of our model to cover such situations is usually rather straightforward. Therefore, our approach is more general and can handle much wider list of practically important situations than just the constant CO input rate design which was selected for illustration and to show how the model and estimation procedure behaves on real data.

We propose to use a formal model simultaneously for CO and radon data which will enable us to estimate both ventilation rate and radon entry rate, in turn. In its formulation, we will start from the very simple, “stylized” theoretical model which has been used in practice of trace-gas experiments previously, Jílek (2003). The theoretical model comprises of a system of two differential equations outlining the CO and Rn dynamic behavior in a closed room after artificial introduction of CO (see the concomitant paper Jílek, Brabec (2004) for detailed technical description of the experiment) in continuous time:

$$\begin{aligned} c'_\tau &= G_C - c_\tau k_\tau \\ r'_\tau &= G_{R,\tau} - r_\tau k_\tau \end{aligned} \tag{1}$$

where:

τ is time (measured e.g. in hours)

c_τ, c'_τ is the CO concentration and its time-derivative, respectively

r_τ, r'_τ is the Rn concentration and its time-derivative, respectively

k_τ is the ventilation rate (assumed to be the same for both tracer gas and radon), which is not necessarily constant in general. Moreover, it is unknown.

G_C is the CO entry rate. It is kept constant in our experiments by an accurate pump, see Jílek, Brabec (2004) for technical details. Its value is known.

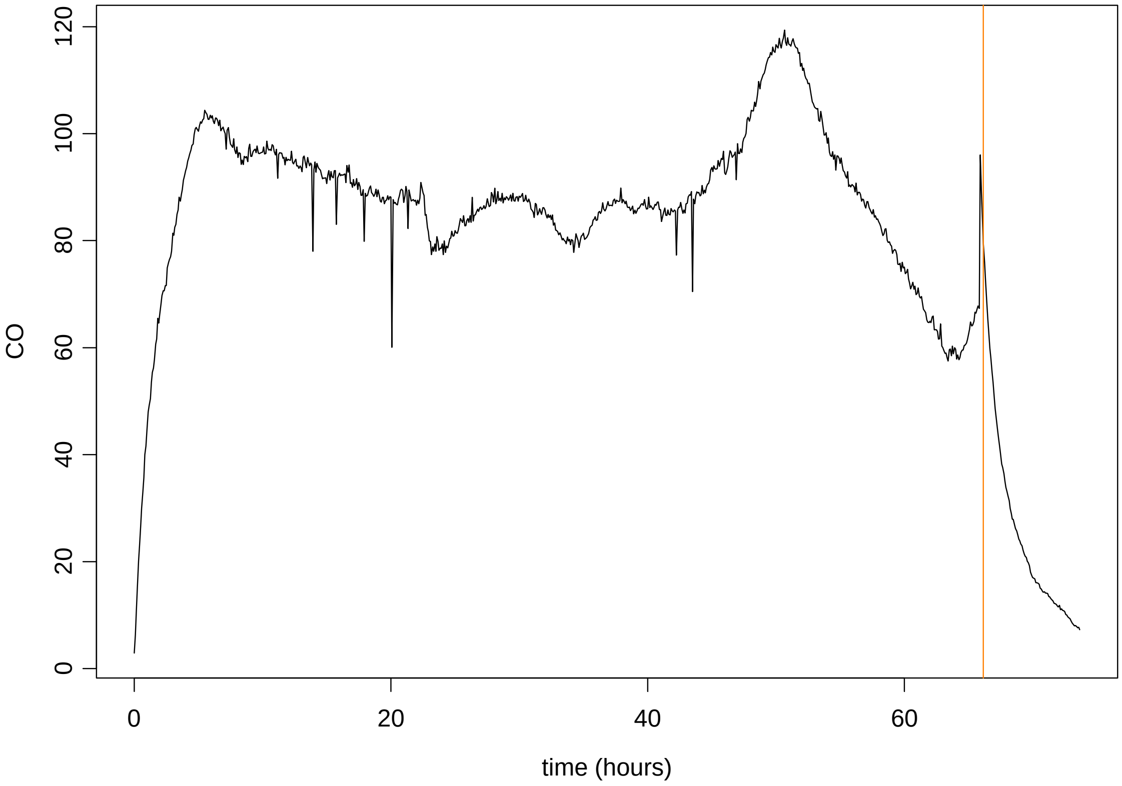
$G_{R,\tau}$ is the Rn entry rate, which does not need to be constant. It is an unknown quantity of ultimate interest.

Note several important features:

- both ventilation rate (k_τ) and radon entry rate ($G_{R,\tau}$) can be time-varying. Obviously, assumptions of constant k and/or G_{Rn} leads to great simplification and easy explicit solution of the system. Such a mathematical convenience should not overwhelm the model construction, however. Namely when dealing with long-term measurements (lasting for several days) in real houses, both k_τ and $G_{R,\tau}$ might change easily. For instance, ventilation rate might change in response to daily temperature changes, among other causes. The concurrent paper Jílek, Brabec (2004) demonstrates on real data, how unjustified simplifying assumptions can easily lead to incorrect estimates. The overall message is clear immediately: model has to respect reality (allow for time-varying rates) and not vice versa.
- ventilation rate k_t is assumed to be the same for both gases, so that it occurs in both equations and interlinks them. This feature is crucial for the tracer gas technique to have any practical relevance. Precisely because of this interlinking, the tracer-gas is able to help us to estimate Rn dynamics. There we do not know two attributes (k_τ and $G_{R,\tau}$), so that one measurement of Rn concentration cannot disentangle them (without heavy assumptions and/or externally supplied knowledge for at least one of them).

Although the equations (1) capture main physical characteristics and essential features of the gas behavior in a closed room, many important details are not specified there (e.g., they do not detail system behavior near origin, when the tracer gas is introduced and when the dynamic situation is clearly more complicated than (1) suggests).

First of all, any practical implementation has to acknowledge discrete time sampling schedule. Moreover, (1) is a theoretical model, and therefore it does not address measurement error presence. In practice, measurement errors might be appreciable even when measurement process is fully under control. Occasional non-negligible errors arising as a result of random electrical shocks to the CO measuring device make the situation even more complicated. Their presence is illustrated on the following figure showing CO measurement readings recorded during a particular measurement campaign (red vertical line shows the time when experimental conditions stopped to be controlled).



Simple minded solution based on plugging in measured concentrations into (1) or its discretized version might lead to serious distortions of concentration dynamics and consequently to incorrect estimates of both ventilation rate and Rn intake rate.

A straightforward solution to the presence of measurement noise would call for pre-smoothing the measured concentrations before using them to solve (1) (or some discrete approximation of it). Concurrent paper Jílek, Brabec (2004) explores various smoothing techniques. Main advantage of the direct smoothing approach is the ease of its use. Some methods are ready for almost pocket calculator use. Main disadvantage stems from the fact that, even each smoothing technique relies on certain statistical and other assumptions, they are not explicitly apparent. Hence, their practical consequences are hard to foresee. Choice of one particular smoothing technique out of many possibilities might be somewhat subjective. Moreover, clear-cut recommendations are difficult to give, since the relative performance of different methods depends on many circumstances (like smoothness of the concentration tracks, signal to measurement noise ratio, sampling interval, etc.). Overall, the use of direct smoothing techniques seems to be somewhat heuristic: it works well in some situations, but the performance is worse occasionally. Theoretically sound recommendation for choice of the method or even for the choice of tuning parameters within a selected method are hard to give in general.

Therefore, we will use another route here. We will formulate a statistical model describing the tracer-gas and radon dynamics during the tracer experiment. Concentration data will be smoothed in a way that arises as a consequence of the model formulation. Moreover, the quantities of main interest: ventilation rate and radon entry rate will come out as one of the model components. The approach has an appealing property: it is "transparent" in the sense that the model formulation and its consequences can be discussed, and if needed, they can be improved in a modular way. The model has several parameters, which govern degree of smoothness showed by the resulting estimates. Unlike in the ad hoc smoothing, these parameters are not selected subjectively by the user, but they are estimated from the data. Maximum likelihood estimation is used here for its generally appealing statistical properties, Schervish (1997).

2. Statistical model formulation and results

In order to have a practically usable tool for real measurement data processing, we will reformulate the "theoretical" model (1) in discrete time as a statistical model, so that we will be able to address non-negligible measurement errors. It is clear that any model serving for this purpose should be dynamic and should achieve some smoothing in the estimates. Ideally, the degree of smoothness should not be chosen arbitrarily, but it should be inferred from the data. One general model specification that satisfies these requirements is the state-space class. This is what we will use subsequently.

2.1.Tracer sub-model

Let us concentrate on the first equation of (1), that is on the equation describing tracer (CO) behavior. We will formulate (a somewhat more detailed) state-space discrete-time analog as:

$$c_t = \gamma_t + \varepsilon_t \tag{2.1.1}$$

$$\begin{aligned} \gamma_t &= \gamma_{t-1} + \Delta.G_C - \Delta.\gamma_{t-1}\left(\kappa_{t-1} + D + g_t\right) + \Delta.\eta_{1t} \\ \kappa_t &= B.\kappa_{t-1} + \eta_{2t} \end{aligned} \tag{2.1.2}$$

$$\varepsilon_t \sim N\left(0, \sigma^2 + p.\gamma_t\right) \tag{2.1.3}$$

$$\begin{pmatrix} \eta_{1t} \\ \eta_{2t} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$$

$$g_t = E.\left(\beta - \alpha.\Delta.t\right).\exp\left(-\alpha.\Delta.t\right) \tag{2.1.4}$$

$$\begin{pmatrix} \gamma_0 \\ \kappa_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad Var\begin{pmatrix} \gamma_0 \\ \kappa_0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & P_{0,22} \end{pmatrix} \tag{2.1.5}$$

where:

Δ is the discrete time step (i.e. the interval between two consecutive tracer measurements). For instance, in the dataset which we use for illustration below, CO was measured every 5 minutes, so that $\Delta = 0.08333$ hour.

$t = 1, 2, \ldots T$ indexes time. Real time τ in hours is obtained as $\tau = t.\Delta$

c_t is the CO concentration measurement reading at time t. In our example, it was measured in ppm.

G_C is the CO entry rate. Since CO was supplied artificially, it is known. $G_C = 44$ ppm/hod was maintained constantly.

γ_t is the true CO concentration.

ε_t is the measurement error. It is exactly the measurement error, what prevents us from seeing the true CO concentration (γ_t) directly. We are able to see only a noisy version of it, namely the c_t. Note that we are allowing for quite general measurement error variance structure. Namely, it has a constant part (roughly interpretable e.g. as "calibration error" variance) and a part which is proportional to the measured value (to the true CO concentration, γ_t), with proportionality constant p. Since there is an ongoing expert dispute about whether the measurement error variance is constant or proportional under practically interesting situations, the formulation is useful both as a precaution and as a tool to address the question formally. When both σ^2 and p parameters are estimated from empirical data, magnitudes of proportional- and non-proportional variability can be compared and their relative importance can be tested formally.

$\kappa_t + D$ is the ventilation rate, which is time-varying, in general. This is the quantity of main interest in the CO measurement endeavor. While it is not directly observable, it is one of the latent model components (state variables), hence it can be obtained relatively easily by appropriate techniques (extended Kalman filter). It comprises of the random (κ_t) and fixed (D) part.

η_{1t}, η_{2t} are two random, normally distributed and uncorrelated structural disturbances, which facilitate dynamic state variables' movements. Their variances govern structural variability. It is clear that larger σ_1^2 and/or σ_2^2 variances mean more movement in the structural part of the model. Their magnitudes have substantial consequences for precision of the κ_t estimates and to their smoothness. When structural variability is large relative to measurement error variability ("signal-to-noise-ratio" is large), model estimates are rather rough. On the other hand, more smoothing is applied when error variability is relatively larger. Hence, it is understandable that it is important to estimate these variances from the data (unless one has substantial a priori knowledge about them) and not to choose them haphazardly (as is the case when using some of the ad hoc direct smoothing methods).

B is the parameter influencing κ_t dynamics and its stationarity/nonstationarity properties. $|B| < 1$ implies stationarity in the κ_t process. That is the tendency to forget initial conditions and to return to zero, and consequently the tendency of the ventilation rate $\kappa_t + E$ to return to a long-term average. Important thing is that the stationarity is not enforced in advance. $B = 1$ is allowed, when data suggest non-stationary behavior. Obviously, when $B = 1$, we are dealing with random walk as an example of non-forgeting dynamics.

g_t is a correction term, which expands scope of the model, as compared with the model (1). g_t term compensates for system's behavior close to the origin (shortly after the artificial CO input started), when the dynamics is more complicated than suggested by the simple physical model (1). Ideally, g_t choice should be based on some physical/chemical theory. After prolonged consultations with measurement experts, no such theoretically derived g_t was available to us, so that we selected $g_t = E.(\beta - \alpha.\Delta.t).\exp(-\alpha.\Delta.t)$ empirically. This is a function that has been used in nonlinear regression problems in the past, Ratkowski (1989). It is a function having one maximum or minimum (depending on parameter signs), for which $\lim_{t\to\infty} g_t = 0$ and which fits starting portions of the empirically estimated ventilation rate tracks quite well under various situations we inspected. Moreover, measurement experts suggest at least some qualitative interpretation to this curve (typically, a curve with minimum is obtained). It relates to the gas expansion in an empty room (which leads to impression of a larger ventilation), combined with temporal "escape" of the tracer into micro-pores, followed by its release (impression of small ventilation) and gradual stabilization around longer term levels. Important thing is that the g_t convergence to zero is very fast, so that the correction is "active" only in the starting part of the ventilation rate track estimate (say less than 2-3 hours from the origin of the experiment), which is not used by traditional methods, Jílek (2003). So that, even if the user would decide not to rely on the compensating function g_t, he/she can still discard the initial part of the ventilation rate estimates in the traditional way and decide not to make use of the extended scope, which the compensation offers. In future, it will be interesting to seek for physically motivated g_t forms. Suggestions from readers are welcomed.

$\begin{pmatrix} \gamma_0 \\ \kappa_0 \end{pmatrix}$ are starting values of state variables γ_t and κ_t. Note that (2.1.5) specification dictates that the true tracer concentration starts exactly from zero (with zero variability). This is to reflect physical arrangement of the experiment.

Such a model specification is rather explicit, compact and modular. (2.1.1) specifies the so called measurement equation. That is, it ties the observable quantity with unobserved states of the system. (2.1.2) specifies state equations, which describe (Markovian) dynamics of the true system states. (2.1.3) describes behavior of the random disturbances, both for measurement error ε_t and for structural disturbances η_{1t}, η_{2t}. (2.1.4) introduces a correction term that empirically compensates for complicated dynamics close to the origin (shortly after the trace starts to be inputted) in order to attempt biased reduction in ventilation rate estimation in that region. To complete the specification, starting values (2.1.5) are specified. Modularity of the specification is convenient. Not only that it helps to understand the model structure, and makes for instance software implementation easy and flexible, but it also facilitates gradual model improvement by future incorporation of additional physical model features (e.g. more detailed physical model of the behavior close to the origin).

Taken as a whole, the dynamical statistical model is formulated as a state-space model, with true tracer concentrations and air ventilation rates as states. From interpretation point of view, the model (2.1) has an interesting property that it disentangles measurement error variability and structural variability driving dynamics of the system. Among other things, it can offer computationally de-noised version of the tracer measurements. More importantly, it can facilitate estimation of the ventilation rate (which is not directly observable and appears in the model as a latent variable only).

From practical point of view, (2.1) can be viewed as a semi-parametric model. It contains some *structural* parameters: $\sigma^2, p, \sigma_1^2, \sigma_2^2, B, D, \alpha, \beta, E, P_{0,22}$ that specify a particular model within larger (and hence flexible) class of models of similar type. Behavior of state components

γ_t , κ_t of utmost interest is specified in an even more flexible way. Their trajectories are not tied by any parametric model, only their smoothness is dictated by quantities analogous to signal-to-noise ratio, which are implied by the structural parameters. Such flexibility is extremely advantageous when no substantial a priori knowledge of γ_t and κ_t is available (other than the measurement experts believe strongly in their smooth, not completely erratic behavior). Wide range of possible trajectories (including trends and near periodicities induced by daily ventilation rate changes related to temperature changes, etc.) When specific physical information becomes available (e.g. for a particular controlled experiment, or when measurement experts provide theoretical physical models and temperature data to incorporate influence of ambient temperature upon the air ventilation rate), more restricted model of γ_t and κ_t tracks should be specified. Such restrictions might improve estimation's precision appreciably (at the cost of smaller generality, of course).

As is usual in the state-space models context, estimation proceeds conveniently via Kalman filtering algorithm. The algorithm goes on recursively, so that even large amounts of data (long-term measurements with short measurement time steps) can be processed easily, Jones (2001). Compared to the classical linear structural time series models, Brockwell and Davis(1991), one additional complication stems from the fact that the model is nonlinear in the states. As a relative easily implemented remedy, we used local linearization, leading to the so called extended Kalman filter algorithm, Harvey (1991).

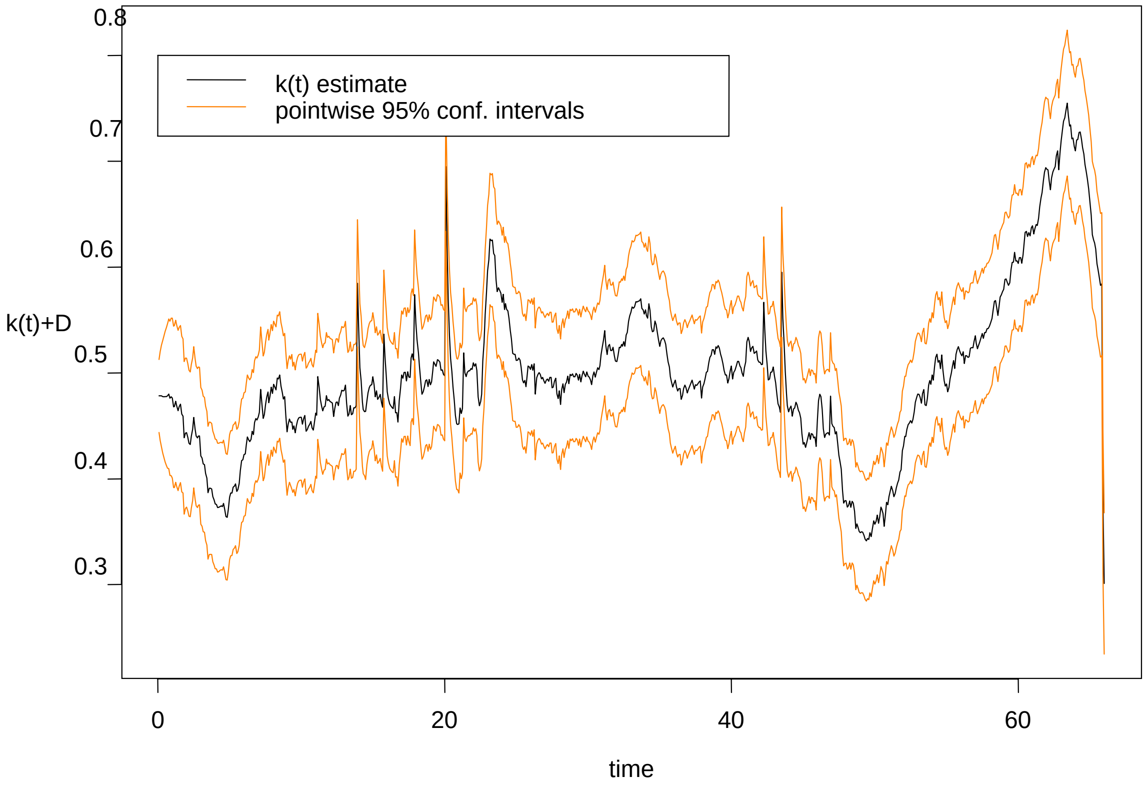
The (extended) Kalman filtering algorithm yields state estimates in the form of the conditional expectation for the quantity of interest, given the past history of various length. The algorithm is advantageous not only from computational point of view, but also because it offers various estimates of state components, which can be useful in different contexts. Here, we will deal with filtered estimates, that is with $\begin{pmatrix} \hat{\gamma}_{t|t} \\ \hat{\kappa}_{t|t} \end{pmatrix} = E\left(\begin{pmatrix} \gamma_t \\ \kappa_t \end{pmatrix} | c_1, c_2, \ldots, c_t\right)$, supplemented with mean squared filtration errors $Var\left(\begin{pmatrix} \hat{\gamma}_t - \gamma_t \\ \hat{\kappa}_t - \kappa_t \end{pmatrix} | c_1, c_2, \ldots, c_t\right)$. For online implementation in terms of a software program giving a preliminary estimates during measurement campaign (useful to check feasibility of the experimental setup for instance), l-step-ahead (l=1, 2, …) predictions $\begin{pmatrix} \hat{\gamma}_{t+l|t} \\ \hat{\kappa}_{t+l|t} \end{pmatrix} = E\left(\begin{pmatrix} \gamma_{t+l} \\ \kappa_{t+l} \end{pmatrix} | c_1, c_2, \ldots, c_t\right)$ might be of interest (together with appropriate mean squared prediction errors). One-step-ahead predictions are obtained automatically as a byproduct of the Kalman algorithm. The predictions for general l can be obtained by a slight extension. In any case, the estimates are nearly optimal from statistical point of view (in the minimum mean square prediction/filtration error sense) when the model holds (they would be exactly optimal if the model were linear).

Another great advantage of the Kalman filtering algorithm is that it can handle missing data and other measurement schedule irregularities quite easily. The algorithm mechanics proceeds alternating one-step-ahead prediction and updating when a new observation is available. When it is missing, all what needs to be altered is that the updating step is omitted and prediction propagates further ahead at the cost of increased state estimate uncertainty (it is normally decreased during information acquisition phase in updating step). Easy handling of missing values is valuable from practical point not only when certain measurements are not taken (skipped for some reason), but also when they their values are highly suspicious and judged as unreasonable by a measurement expert. Our suggested robust measurement data processing should go in two steps: i) expert inspection of the measurement series and exclusion of data with obvious errors and/or data highly suspected as erroneous. The excluded data points are then flagged as missing, ii) resulting series is processed via Kalman filtering. In practice, we have applied the robustified procedure when analyzing the real measurement data shown in the section 1 (results are shown below as an illustrative example). There, we had to exclude (and take as missing) four data points with unrealistically low CO readings which are believed to be caused by electric shocks to the CO sensor. The real impetus of this procedure lies in the fact that highly suspect measurement values are not let into the estimation procedure so that they do not ruin structural parameters estimates. On the other hand, the suspicious values are not replaced by any fabricated value. Data processing proceeds as if they are missing, hence it does not introduce any systematic bias (as long as the model (2.1) is roughly correct), their absence is reflected in increased uncertainty in the estimates.

During the course of computations, the Kalman algorithm produces not only the point estimates of the quantities of interest, but also estimates of their filtration/prediction variability. This is highly desirable, because they can be used (together with the assumed normality of measurement and structural errors) to produce pointwise approximate (i.e. asymptotic) 95% confidence intervals according to the simple formula: (point estimate) ± 1.96*(estimation standard error). It is important to note that such an interval has approximately the nominal coverage at a particular time point. These intervals are not constructed to keep the coverage simultaneously. In particular, the band given by lower and upper limits taken as functions of time is not guaranteed to capture true track with the nominal confidence.

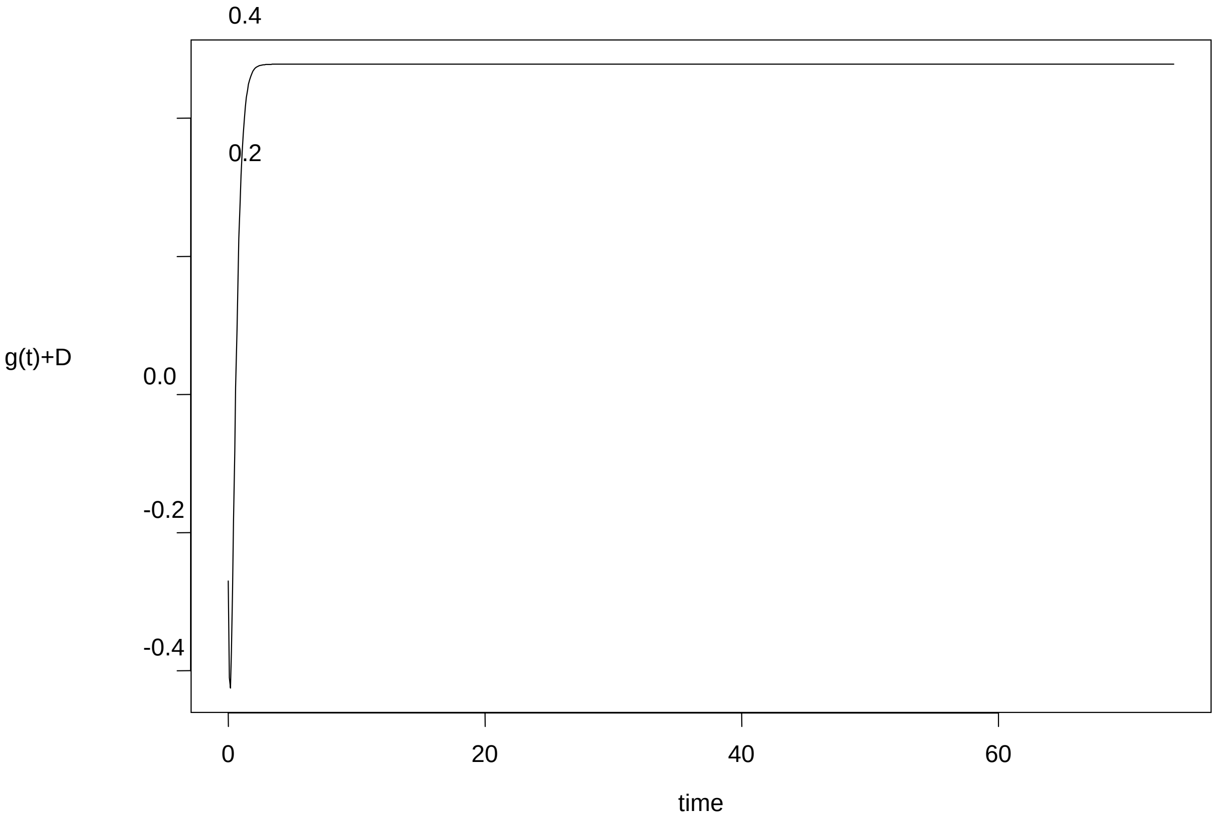
As specified so far, the model (2.1) and subsequent Kalman filtration are not directly usable for data processing because several *structural* parameters $\underline{\theta} = \left(\sigma^2, p, \sigma_1^2, \sigma_2^2, B, D, \alpha, \beta, E, P_{0,22}\right)'$ which appear in its formulation are generally unknown. Hence, they have to be estimated from data. We will use maximum likelihood estimation for its appealing asymptotic properties, Schervish (1997) and rather easy implementation. Likelihood for state space models can be evaluated rather easily via prediction error decomposition, Brockwell and Davis (1991). The decomposition is facilitated by Kalman filtering algorithm output in a computationally manageable way. Maximization of the log-likelihood with respect to $\underline{\theta}$ gives the maximum likelihood estimate (MLE) of structural parameters. This has to be carried out numerically, nevertheless the estimation procedure can be programmed quite easily. We used S-plus (version 6.1) programming and statistical package environment, Chambers, Hastie (1992). Important thing is that the estimation has to be carried out only to get structural parameters values in order to be able to construct the Kalman filter (that is to “train” the Kalman filter). Practical implementation of the filter to future data from similar experimental setups might use the $\hat{\underline{\theta}}$ estimate obtained earlier. From operational point of view, it means that possible practical software implementation might include the structural parameter estimation procedure (which is more demanding computationally, but offers more generality of the resulting software), or it might go without it. Then it would have to take externally estimated $\hat{\underline{\theta}}$’s (e.g. from specialized statistical software) or take $\underline{\theta}$ as fixed for similar experimental setups (such implementation would be much easier to program, at the cost of somewhat sub-optimal statistical properties of the filter).

The next plot shows estimates of the air ventilation rate computed via extended Kalman filter after estimating structural parameters via maximum likelihood from a data set obtained during a measurement campaign conducted in a single room of a real house. Time is measured in hours. Point estimates of ventilation rate are supplemented with 95% confidence intervals. The confidence intervals are asymptotic (based on local linearization) and they are constructed pointwise.



From there, one can see that the estimated air ventilation rate trajectory is quite complicated and is rather far away from being constant, although in the substantial part of the measurement period, it wanders roughly around 0.47, which is the MLE estimate of the D parameter.

The next plots shows the shape of the estimated function $g_t + D$ (MLE obtained from the same data). It illustrates the qualitative features of the near-origin-compensator function behavior, which are typical for other measurement campaign data as well.



This illustrates transient nature of the g_t compensator. It stays active only close to the origin, later on, it quickly converges to zero (after going through a minimum), so that $g_t + D$ goes to "long term" κ_t value.

Maximum likelihood estimates of the structural parameters of the model (2.1) are summarized in the following table. Loglikelihood evaluation at the MLE estimate gives -270.074

Structural parameter	Maximum likelihood estimate
σ^2	1.56716
p	0.00000
σ_1^2	0.17006
σ_2^2	0.00016
B	0.98281
D	0.47811
α	3.65523
β	-0.49940
E	1.49858
$P_{0,22}$	0.00000

Not only that these estimates are inputs to the full specification of the Kalman filter and that they are obtained via formally justified estimation procedure form data (unlike more or less subjective choices that has to be made with direct ad hoc pre-smoothing of CO measurements), but their careful inspection can give us various valuable insights about the physical behavior of the measured system and its dynamics. For instance, the B value 0.98 is smaller than 1 in absolute value, suggesting stationary behavior of the ventilation rate process around the long term value D=0.48. Since the B value is very close to one, the process has "long memory" and resembles a random walk to a large extent. This means that the departures from the long term value can be substantial and they can persist for non-negligible periods of time. Inspection of measurement error variability related parameters turns out to be interesting as well. The estimates suggest that the heteroscedastic measurement error variability (whose variance is proportional to the measured value) is rather negligible ($\bar{p}$ =0) and hence that the measurement error variability is predominantly non-proportional (homoscedastic), suggesting perhaps that the major part of the measurement error variability might come from the calibration error.

There are important practical situations when the analysis done so far would be all what is needed. Civil engineering applications when only the air ventilation rate for a room is to be estimated can serve as an example. On the other hand, when the goal is to estimate radon entry rate, air ventilation rate estimated so far is not of direct interest and plays a role of a nuisance parameter which has to be estimated in order to be able to estimate Rn entry rate. To complete that, one has to specify the radon sub-model (a discrete time analog of the second equation of the model (1)). It will be done in the next section.

2.2. Radon sub-model

Let us proceed with the Rn sub-model, that is with the discrete time analog of the second equation of (1):

$$r_{t'} = \rho_{t'} + \psi_{t'} \tag{2.2.1}$$

$$\begin{aligned} \rho_{t'} &= \rho_{t'-1} + \Delta_R . G_{t'-1} - \Delta_R . \rho_{t'-1} . (\kappa_{t'-1} + D) - \Delta_R . \rho_{t'-1} . \omega_{1t''} \\ G_{t'} &= G_{t'-1} + \omega_{2t'} \end{aligned} \tag{2.2.2}$$

$$\psi_{t'} \sim N\left(0, \sigma_R^2 + p_R . \rho_{t'}\right) \tag{2.2.3}$$

$$\begin{pmatrix} \omega_{1t'} \\ \omega_{2t'} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Delta_R^2 . \sigma_{k,t'}^2 & 0 \\ 0 & \sigma_{R,2}^2 \end{pmatrix} \right)$$

$$\begin{pmatrix} \rho_{0'} \\ G_{0'} \end{pmatrix} = \begin{pmatrix} a_{01} \\ a_{02} \end{pmatrix}, \qquad Var\begin{pmatrix} \rho_{0'} \\ G_{0'} \end{pmatrix} = \begin{pmatrix} P_{R0,11} & 0 \\ 0 & P_{R0,22} \end{pmatrix} \tag{2.2.4}$$

where we have (in addition to quantities already described in the previous subsection):

Δ_R is time step of the radon measurements. In our illustrative real data example, we have the radon measurements taken 1 hour apart, so that Δ_R=1.

t' is time index of the radon measurements. Actual time (in hour from the beginning of the radon measurements) is given as $\tau = t'.\Delta_R$. The elaborate t' notation is used to stress that the radon time index might be different from CO time index t (in our example, CO measurements were taken each 5 minutes, while Rn readings were available only each hour, so that $t = 12.t'$).

$r_{t'}$ is the Rn measurement reading (in Bq/m^3). It measures average Rn activity taken in the time interval $(t'-1, t']$.

$\rho_{t'}$ is the true Rn average activity in the time interval $(t'-1, t']$.

$\kappa_{t'} + D$ is the ventilation rate, as it appears in the tracer sub-model (2.1), except for the fact that it has to be averaged in the $(t'-1, t']$ interval in order to match the $r_{t'}$ and $\rho_{t'}$ timing.

$\psi_{t'}$ is the measurement error. As can be seen from (2.2.3), it has generally heteroscedastic variance with one part proportional (with proportionality constant p_R) to the measured value ($\rho_{t'}$) and the other constant (σ_R^2), independent of the measured quantity.

$\omega_{1t'}, \omega_{2t'}$ are independent structural disturbances which are driving stochastic dynamics other than that coming from $\kappa_{t'}$ movements. Their covariance matrix is now time-dependent in order to reflect uncertainty coming from the fact that the $\kappa_{t'}$ is not known but estimated from the tracer data and the tracer sub-model. See the text below for the motivation.

$G_{t'}$ is the radon entry rate (in Bq/(m^3.hod)). This is the quantity, whose estimate can be viewed as the ultimate goal of the analysis comprising of the modeling process and subsequent estimation. It is generally time-varying. The model does not constrain it to follow any concrete parametric trajectory. In particular, it is not enforced to be constant. On the other hand, if the data are not inconsistent with constant $G_{t'} = G_R$, we will be able to note that from the results. In this sense, the model represents a valuable explorative tool which can help us to learn about the qualitative features of the Rn entry dynamics in real-world conditions without potentially problematic heavy a priori assumptions.

$\sigma_{k,t'}^2$ is the variability of the prediction/filtration error in $\hat{\kappa}_{t'}$ estimates obtained from tracer sub-model (2.1). This variance describes uncertainty in the $\hat{\kappa}_{t'}$ estimates and hence it is different for different t' (i.e. time-dependent).

The radon submodel (2.2) is specified as a state-space model in discrete time. (2.2.1) is the measurement equation. (2.2.2) represents the (simple Markovian) state equations. (2.2.3) specifies random disturbances behavior (they are normally distributes and mutually independent). (2.2.3) specifies initial conditions which are unknown and has to be estimated from the data (due to the nonstationary, random walk character of the $G_{t'}$ process, the initial value matters and cannot be set to arbitrary value in the hope that its influence will diminish in time).

Structure of the model comes essentially as an analogue to the “smooth trend” model used heavily in econometrics, Koopman et al (1995), albeit with more complicated (time-varying) transition matrix. Such a model is selected to reflect assumption of substantial smoothness in the radon entry rate dynamics (measurement experts believe that large, short term fluctuation of the entry rate are extremely unlikely). This is a model that we get when we think of $(\kappa_{t'} + D)_{t'=1,\ldots T'}$ as of a known sequence. That would give exactly (2.2) except for the first equation of (2.2) where the last term would be missing and for (2.2.3) where $\omega_{1t'}$ term would be missing as well. To acknowledge variability in the $\hat{\kappa}_{t'}$ estimates plugged into the “essentially smooth trend” model, one can take their estimation errors ($\sigma_{k,t'}^2$) into account and apply local linearization of the model. Using approximate unbiasedness property of $\hat{\kappa}_{t'}$, we arrive at the formulation (2.2). Altogether, (2.2) differs from the “essentially smooth trend model" in recognizing uncertainty in $\hat{\kappa}_{t'}$ plug-ins in. Note that this acknowledgement does not increase the number of structural parameters (since the approximate increase in the structural equations variability is completely given by $\hat{\kappa}_{t'}$'s uncertainty).

Model structural parameters: $\underline{\theta}'_R = \left(\sigma_R^2, p_R, \sigma_{R,2}^2, a_{01}, a_{02}, P_{R0,11}, P_{R0,22}\right)$ are estimated by maximum likelihood estimation procedure. It uses prediction error decomposition idea in exactly the same way as before. Once the structural parameter values are available, the extended Kalman filter can be run easily to get filtered estimates and/or predictions of the state variables $G_{t'}$ and $\rho_{t'}$.

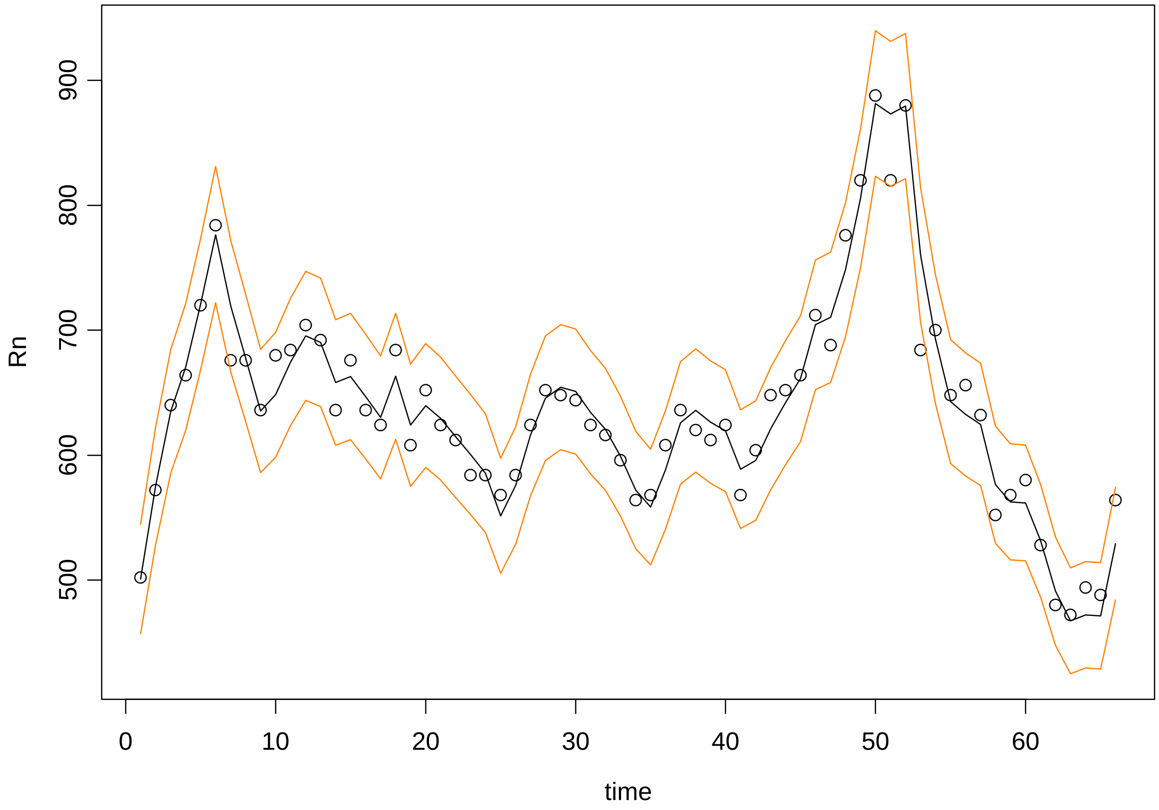
While the $\hat{\rho}_{t'}$ estimate is of secondary interest (it can be used to smooth or de-noise Rn measurement computationally), obtaining the $\hat{G}_{t'}$ estimate represents the ultimate goal of the whole analysis. $\hat{G}_{t'}$ is the estimate of the (possibly time-varying) radon entry rate. Great advantage of the dynamic state space mode (2) formulation is that it produces not only the point estimate for each time within the observed period (including the times when observations were not taken), but it provides their standard errors as a byproduct. Hence, approximate 95% confidence intervals can be easily constructed in pointwise fashion.

Let us look at some of the model results, obtained from real data acquired during a particular measurement campaign in a single room of a selected house (the same campaign, whose data were used for tracer calculations and air ventilation estimation in the previous subsection). Following table shows maximum likelihood estimates of the structural parameters. Loglikelihood evaluated at the maximum likelihood estimate gives -539.726

Structural parameter	Maximum likelihood estimate
σ_R^2	0.0000
p_R	130.0902
$\sigma_{R,2}^2$	0.5644
a_{01}	352.8545
a_{02}	316.5501
$P_{R0,11}$	0.0000
$P_{R0,22}$	0.0000

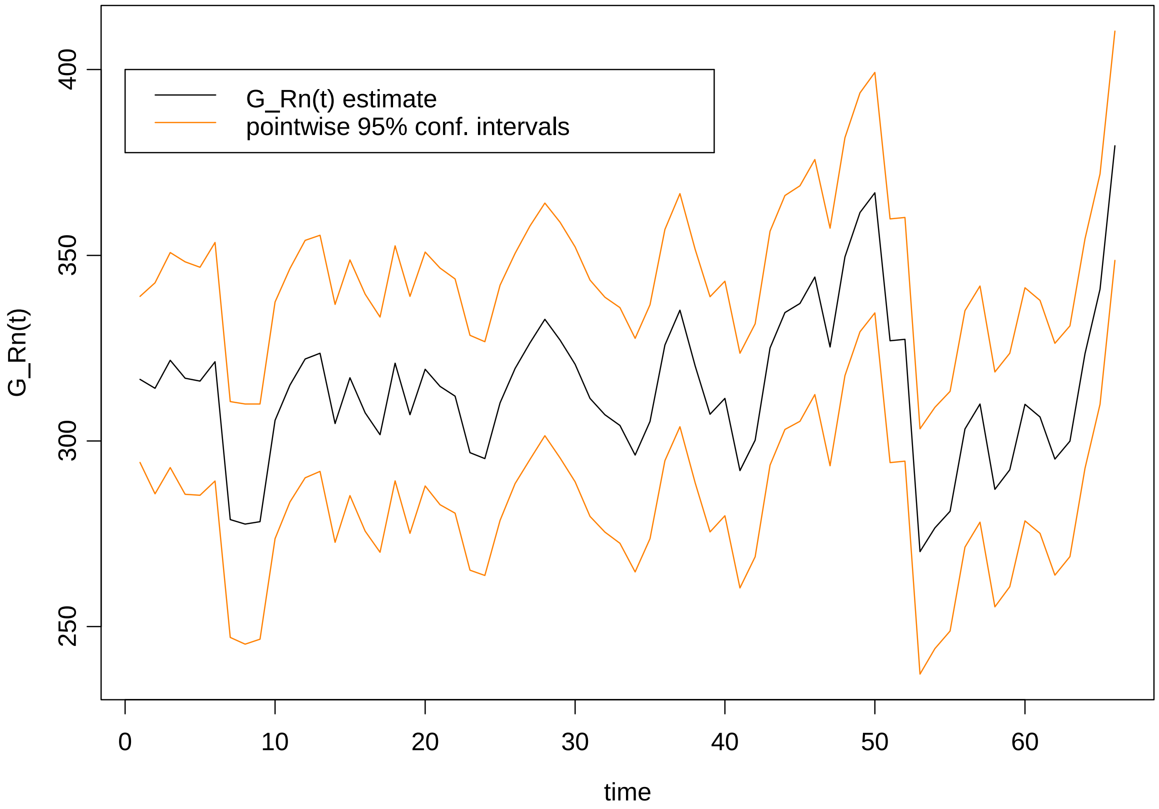
It is interesting to think about interpretation and physical consequences of actual parameter values. One of the advantages of the state-space model specification is the fact that many model features, including structural parameters have directly interpretable physical meaning. For instance, zero value of σ_R^2 suggests that the measurement error is purely heteroscedastic, its variance depends approximately proportionally on the measured value. This is not surprising, when thinking about approximate Poisson behavior of the measured counts. Initial value of radon entry rate was estimated as a_{02}=316.5501. Since the $G_{t'}$ process is a random walk with non-negligible innovation variance, the radon entry rate wanders around this value in quite erratic way.

Even though the model (2) is flexible, it enforces some structure in the data, due to its assumptions about the data-generating process. Hence, it is of interest to verify how it fits empirical data. For the radon sub-model specifically, it is interesting to check whether the behavior of the model filtration of true radon activity (i.e. $\hat{\rho}_{t'}$ estimates) is not inconsistent with the behavior of the actual Rn measurements. Following plot compares actual measurements (dots) with $\hat{\rho}_{t'}$ based on the model (2) (black solid line) and their pointwise 95% confidence interval limits (red solid lines).



From there, we can see that the model fits the data well. No evidence of any substantial misfit is found here. Note that the $\hat{\rho}_{t'}$ estimates are able to follow complicated trajectory of actual measurements quite closely. Estimate of the $\hat{\rho}_{t'}$-estimation error (which is used in confidence intervals construction) seems to be sufficient to cover estimate's variability (perhaps even a little bit too conservative, occasionally).

The next plot shows radon entry rate estimates, $\hat{G}_{t'}$ as a function of real time. This is the actual goal of the whole analysis. Point estimates are supplemented with pointwise constructed 95% confidence intervals.



We can see that even though the estimated radon entry rate wanders a little bit, the confidence intervals do not suggests that it should change dramatically. Note that even if the pointwise confidence intervals do not have the nominal coverage simultaneously, the conclusion is on the safe side, because the simultaneous intervals should be actually even wider. Hence, if we are willing to accept the conclusion that the radon entry rate G_t did not change during the measurement period, then it might be sensible to think of the mean Rn rate. Arithmetic mean of the estimates comes out as 314 Bq/hod. Incidentally, it is very close to the estimated starting value of 316.55 Bq/hod, which once again suggests that constant entry rate seems plausible and that the observed variability of radon activity measurements comes predominantly from measurement error and air ventilation rate variability (which seems to vary during the campaign, see section 2.1 results).

3. Conclusions

In this paper, we formulated a dynamic statistical model that is suitable for modeling data obtained from tracer experiments which include tracer gas (e.g. CO) and radon measurements in time series. Its formulation starts from a simple theoretical physical model (1), but then the model is made more concrete and practically usable by explicit acknowledgement of two major complications: i) discrete sampling of the underlying continuous process, ii) presence of the measurement error. Details of our model specification capture various features not present in the physical ideal model (1). Those include: quite precise specification of the measurement error statistical behavior, correction for the complicated dynamic behavior at the beginning of the tracer experiment, detailed description of the system dynamics by Markovian state transition structure.

The statistical model is specified as a dynamic state-space nonlinear model (2). Actual parametrization is selected in such a way that both state components and structural parameters have a physical meaning and definite interpretation. The model comprises of two sub-models (tracer and radon submodels, (2.1) and (2.2) respectively), which are tight together by the same air ventilation rate. Although the model works with bivariate measurements (CO and Rn) in principle, we decouple the estimation into two phases: i) estimation of air ventilation rate from tracer (CO) data, ii) estimation of radon entry rate using radon measurements and the air ventilation rate obtained in the previous step. This two-phase estimation was preferred to fully bivariate observation model for three reasons: i) it simplifies implementation and ii) it prevents Rn back-influence upon air ventilation rate estimates (which would occur in the bivariate approach and might introduce problems since the Rn measurements are taken on much coarser time scale), iii) it is able to use the tracer sub-model separately if needed (e.g. in civil engineering applications when only the ventilation rate is to be estimated via tracer measurements and no Rn measurements are taken).

Estimation of quantities of interest (air ventilation rate and radon entry rate, which appear as state variables in the model) proceeds via extended Kalman filter algorithm. Customarily, it uses local linearization of the originally nonlinear model. Not only that the algorithm provides point estimates of latent components, but is also produces their standard errors as byproducts. They can be easily combined to produce pointwise confidence intervals. Reporting confidence intervals is important from practical point of view: they provide an easily communicable idea of precision.

The model (2) has several structural parameters, whose values are typically not known a priori. Unlike in the ad hoc methods of pre-smoothing, they are estimated from data and not selected more or less subjectively by the user. We use maximum likelihood estimation, which is known to have many appealing statistical properties. This is important since some of the structural parameters (or their ratios) determine degree of smoothness implied in the estimated ventilation rate and radon entry rate time tracks. This smoothness is tuned up in response to what the data say and not subjectively. Consequently, the approach is rather flexible and should be able to handle various situations (e.g those with different air ventilation rate and with different amount of dynamics of the rate) quite easily (and in a way that is essentially optimal from statistical point of view).

Another practical advantage our model (2) offers is the fact that it is specified in a modular way. This means easier interpretability, maintainability and flexibility. For instance, when more detailed physical knowledge becomes available to model tracer behavior close to the origin, one can attempt to build the additional physical formalization into the model (2.1) to improve it. Furthermore, only minor changes in some of the model components are required to cover other tracer experiment designs (e.g. designs when decline of tracer concentration is measured after saturating the measured room by the tracer and stopping its input, with which we dealt in the past, or even more complicated designs with repeated saturation and depletion to address questions about memory effects, etc.).

The model is rather easy to implement in software form. Recursive nature of the Kalman algorithm allows for easy application even for massive datasets (short time steps, longer measurement periods). Application can be easily run in an online mode and might even include short time prediction (e.g. for quick checks of the experimental setup helpful to a technician running the measuring devices). The algorithm can easily handle missing data and measurement schedule irregularities. This is beneficial since some measurements might not be available and because this feature can be utilized for a simple robustifying procedure:

highly suspicious values would be flagged as missing by a measurement expert, and the resulting series with missing values could be easily processed without difficulties. This is because the Kalman filter proceeds in two subsequent steps: i) one-step-ahead prediction, ii) updating once additional measurement becomes available. Missing datum leads essentially only to exclusion of the updating step (followed by uncertainty increase to reflect the missing information).

As we demonstrated on real measurement campaign data (measurement of a single closed room in a real house), the suggested model and subsequent procedure works easily on field data. Obtained air ventilation estimates suggest that assumptions of constant ventilation rate (which had been adopted for convenient estimation in the past occasionally) might not be reasonable at all and hence the more flexible models like the presented one should be used. Radon entry rate estimation procedure gave results comparable to values expected by experts. Unlike the air ventilation rate, the radon entry rate does not seem to be varying substantially during the measurement period in our illustrative example. Close inspection of structural parameters which have direct physical meaning gives interesting insights into the measurement process and statistical features of both tracer and radon dynamic behavior. For instance, the measurement error variance is proportional to the measured radon value (not surprisingly when one thinks of Poisson nature of measured counts). On the other hand, it has only non-proportional components for CO measurement.

4. Literature

Brockwell, P.J.-Davis, R.A. (1991): Time series analysis: Theory and methods. Springer. New York.

Chambers, J.M.-Hastie, T.J. (eds.) (1992): Statistical models in S. Wadsworth and Brooks/Cole, Pacific Grove, California.

Harvey,A.C. (1991): Forecasting, structural time series models and the Kalman filter. Cambridge University Press. Cambridge.

Jílek,K (2003): Continuous monitor of CO gas for determination of ventilation rate, Research report in framework of Institutional Research , National Radiation Protection Institute , Praha. (in Czech)

Jílek,K.–Brabec,M. (2004): Radon diagnostics and tracer gas measurements. 4th European Conference on Protection against radon at home and at work. Praha.

Jones,R.H. (2001): Longitudinal data with serial correlation: a state-space approach. Chapman and Hall/CRC. New York.

Koopman,S.J.-Harvey,A.C.-Doornik,J.A.-Shephard,N.(1995): Stamp. Structural time series analyzer, modeler and predictor. Timberlake. London.

Ratkowsky,D.A. (1989): Handbook of nonlinear regression models. Marcel Dekker. New York.

Schervish, M.J. (1997): Theory of statistics. Springer New York.