AN EXPLORATION OF DATA DRIVEN MODELS FOR ODOR ASSESSMENT

Extended abstract

Mauri Favaron¹,²*, S. Cesco¹, Daniele Fraternali², Gianluigi Valli¹, Roberta Vecchi¹

¹Università degli Studi di Milano, Dipartimento di Fisica, Via Celoria 16, Milan, Italy
²Servizi Territorio srl – Cinisello Balsamo, Via Garibaldi 21, Milan, Italy

Background

Odor perception, occurring in correspondence of single inspiration acts, is characterized by a time scale in the order of 12 seconds (Schauberger et al., 2012). Ideally, modeling odor concentration should then be made on a comparable time frame. The concentration of odor-carrying compounds may change significantly over such a short period, often exhibiting intermittency (Hildeman and Wilson, 1999) and other complex features (Nicell, 2009).

Conventional dispersion models and the meteorological data feeding them are not typically able to resolve so a short range, their long term (most often hourly) averaging actually filtering most interesting variation out (Nicell, 2009; Schauberger et al., 2012). To circumvent this problem various strategies have been developed, as for example using probabilistic models allowing estimation of concentration distribution instead of just hourly means (e.g. SciPuff, online at http://www.sage-mgt.net/services/modeling-and-simulation/scipuff-dispersion-model, and coupled particles models), or adopting an estimate of maximum concentration from its hourly value using peak-to-mean ratio (Nicell, 2009; Schauberger et al., 2012).

In this work an alternative model, developed by the authors, has been explored. It is based on a modified version of Lagrangian particle models in which frequently updated measured statistics replace estimates based on similarity theory. Further work is in progress and will be addressed in a more extensive paper, currently under preparation. In conventional Lagrangian particle models like Flexpart (Stohl et al., 2005) wind statistics are modeled by particular solutions of Langevin-type equations, usually derived under the assumption of a time scale significantly longer than typical Lagrangian decorrelation time (e.g. hourly). These solutions, characterized by quite a simple analytical form, are well suited to estimating hourly concentration, but tend to not provide detailed information about instantaneous concentrations.

Therefore a need exists of a different approach to modeling, in which short time frame is explicitly considered. In this exploration, a particle model has been used in which Langevin-type equation has been replaced with empirical multivariate distributions obtained from high-frequency 3D wind measurements by a three-axial ultrasonic anemometer. This approach, suitable for efficient implementation, has the potential of real-time operation on small-scale, easy to deploy computing devices. The new class of models built this way has been provisionally named of data-driven models. The specific data-driven model used in our exploration is named NanoPart.

Methods

The general Langevin equation for transport and diffusion is (Eq 1):

\[
dv_i = a_i(x, v_i, t) \cdot dt + b_i(x, v_i, t) \cdot dW_i
\]

(1)
where: \( x, v \) are the position and velocity, respectively, and \( dW_j, j = 1,2,3 \) are the components of a Wiener process with zero mean and unit standard deviation.

Eq.1, of generic type, is not directly suitable for use in particle dispersion models, and is used as a basis to derive application-specific analytical forms. For example, in FLEXPART eq.1 is simplified to a set of three independent equations for \( du, dv, dw \). The equation for vertical wind component is (Eq. 2):

\[
dw = -w \frac{d}{dt} \frac{\sigma^2_w}{\tau_{w,z}} dt + \frac{\sigma^2_v}{\rho} \frac{d}{dt} dt + \left( \frac{2}{\tau_{w,z}} \right)^{1/2} \sigma_w dW
\]

where: \( dW \) is a Wiener process of mean 0, \( \tau_{w,z} \) is the Lagrangian decorrelation time along \( z \) axis, \( \rho \) is air density, \( \sigma_w \) the standard deviation of vertical wind component. Eq. 2 represents a special case of Eq. 1, in which all the involved quantities can be directly measured or estimated through similarity theory. This allows using (Eq. 2) (and similar relations for horizontal components) as a device for generating random wind values whose values can then be imparted to pollutant “particles”, simulating transport and turbulent diffusion.

Equations like (Eq. 2), derived under the assumption \( dt \gg \tau_{w,z} \), imply that probability distribution is normal for all wind components, and wind components are pair-wise uncorrelated. These two statements are approximations to reality, valid if a scale much longer than Lagrangian decorrelation time is used. If \( dt \leq \tau_{w,z} \), then neither normality nor lack of correlation hold, an example on real data is given in Fig. 1.a and 1.b.

Fig. 1. Example of wind non-normality and correlation in short-term (20s) time frame

In data-driven models Eq1 is replaced by (eq. 3)

\[
x_p(t + dt) = x_p(t) + V(t) \cdot dt
\]

where: \( x_p(t) \) represents particle position at time \( t \), and \( V(t) \) is a random variable whose statistical distribution (of unspecified analytical form) coincides with the multivariate distribution of wind components as determined by experimental measurements. To apply Eq. (3) in a data-driven model, random values should be generated from the instantaneous distribution of wind components, \( F(v, t) \). As this distribution is not constricted to a specific analytical form, its estimation \( \hat{F}(v, t) \) is made from the set of wind data sampled from \( t - \Delta t \) to \( t \) (Eq.4),

\[
S_{\Delta} = \{ v_j \mid -\Delta t \leq t_0 + k \cdot \Delta t \leq t \}
\]

where: \( \Delta t \) is a user-selected time sequence duration (typically in the order of \( \tau_{w,z} \)), \( \Delta t \) is the anemometer’s sampling rate, and \( t_0 \) a reference time.

In principle the estimate may be computed in different ways. For example, the joint frequency distribution could be approximated using a continuous function (like for example a “gamma estimate” or a combination of B-splines), and random values \( v_j \) extracted from it by means of one of the many existing algorithms. In this work, an implicit method is preferred, in which random values \( v_j \) are obtained by sampling with repetition from the measurements data belonging to \( S_{\Delta}(t) \). If the number of elementary wind data in \( S_{\Delta}(t) \), \( N = \frac{\Delta t}{\Delta t} \), and the number
of random values are sufficiently large so that the joint frequency function of the values approximates the original wind distribution, at a fraction of the computational cost. The ability to reproduce the short time-scale distribution of wind, including its non-normality and correlation structure, without heavy assumptions permits data-driven models to address situations like odor assessment, accidental toxic releases, defense and civil protection.

In analogy to conventional particle models, in data-driven models the wind values \( \mathbf{v}_j \) are fed into (Eq. 3) and use relation (Eq. 5) to update particle positions:

\[
x_{j+1}^i = x_j^i + \mathbf{v}_j \cdot \tau
\]

where index \( j \) spans the set of particles existing at time \( t_i = t_0 + (i-1) \cdot \tau \), and \( \tau \) is particle release time. If \( \tau \) is small enough, many concentration “snapshots” can be computed on each hour, in number sufficient to compute both mean and standard deviation, and potentially to determine point-wise non-parametric estimates of concentration distribution.

Finite particle life time is modeled through the combination of a maximum particle time span \( T \) and an exponential decay function \( M(a) = M_0 \exp(-\beta a) \), being \( a \) the particle age, and different values of \( \beta \) coefficient account for different chemical species. The final model output is a sequence of ground concentration fields. These may be estimated in two ways, selected by the user. The first method counts the \( N_p \) particles contained in a box with basis centered at a receptor \( \mathbf{x} \) and with edges of length \( \Delta x \), \( \Delta y \) and \( \Delta z \). Concentration is expressed as (Eq. 6):

\[
C(x) = \frac{1}{\Delta x \Delta y \Delta z} \sum_{i=1}^{N_p} M_i(a_i)
\]

where: \( M_i(a_i) \) represents the mass of \( k \)-th particle, and the sum is restricted to particles belonging to the box. In the second method each particle is represented as a Gaussian distribution

\[
C_p(x) = \frac{M_i(a_i)}{\sqrt{(2\pi)^3 \sigma_x \sigma_y \sigma_z}} \exp \left\{ -\frac{\left( (x_p - x)^2 + (y_p - y)^2 + (z_p - z)^2 \right)}{2\sigma_x^2} \right\}
\]

where: \( x_p \) represents the particle position, and \( \sigma_x, \sigma_y, \sigma_z \) are estimated from measured wind component standard deviations and Lagrangian decorrelation times. Final concentration at a point is then estimated as

\[
C(x) = \sum_p C_p(x)
\]

where: the sum extends to the whole particle set.

**Results and discussion**

Fig. 2 shows the distribution of particles obtained from running the NanoPart model on a test case characterized by wind meandering. The particle survival time has been fixed to 7200s, with 10 particles emitted every second from the only point source placed at origin with a zero initial speed and at ambient temperature, so that buoyancy can be excluded; \( \Delta t \) was assumed to be 20s, and ultrasonic anemometer data rate set to 10Hz.

![Fig. 2. Example of NanoPart particle distribution, at instant 03:31:45 from simulation start](image-url)
Respect to the cloud of points typically visible using conventional particle models, Fig. 2 shows a defined plume advecting and widening due to turbulence, with meandering effects clearly visible. Other advantages of data-driven models include extreme simplicity (the Fortran 2003 code used to generate the data presented in Fig. 2 is 609 lines long, comments and spaces included), computational efficiency (especially if implicit estimation of wind joint frequency function is used), and possibility of real-time implementation.

Concluding remarks

For a Fig. 2 like to be meaningful, wind measurements in $S_\omega(t)$ must contain detailed information about turbulent fluctuations. This virtually rules out mechanical anemometers, whose inertial low-pass filtering and vulnerability to wind calms cut essential information away. Ultrasonic anemometers or other fast and accurate advanced sensors have then to be used to feed data-driven models. As the vertical wind must be considered, too, three-dimensional instruments are needed.

In addition to the need of advanced anemometers, data-driven models have shown some points susceptible of possible improvement. Among them are worth mentioning the modeling of wind distribution change with position, and the possibility of an optimal selection of $\Delta t$.

Next planned developments include validation campaigns made using non-reactive tracers, mean and standard deviation comparison with prediction of existing probabilistic dispersion models, and the realization of an integrated data acquisition and real time modeling for field use.

Keywords: dispersion models, odor assessment, pollutant instantaneous concentrations, ultrasonic anemometer

Acknowledgements

The authors thank warmly the Meteorological Office of Regional Environmental Agency of Lombardia (ARPA Lombardia), which provided vast samples of ultrasonic anemometer data from their SHAKEUP network, and the SonicLib working group, who processed the data and transformed them for an easy reading of ultrasonic anemometer data. SonicLib, an R library for processing ultrasonic anemometer data, is online at http://www.boundarylayer.eu/projects_soniclib.html

References

Hildeman T.L., Wilson D.J., (1999), Simulating concentration fluctuation time series with intermittent zero periods and level dependent derivatives, Boundary Layer Meteorology, 91, 451-482
Nicell J.A., (2009), Assessment and regulation of odour impacts, Atmospheric Environment, 43, 196-206