Journal of Hydrology 535 (2016) 688–698

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Including land use information for the spatial estimation of groundwater quality parameters – 1. Local estimation based on neighbourhood composition

T. Heißerer [a], C.P. Haslauer [b,*], A. Bárdossy [a]

[a] University of Stuttgart, Institute for Modelling Hydraulic and Environmental Systems, Department of Hydrology and Geohydrology, Pfaffenwaldring 61, 70569 Stuttgart, Germany
[b] University of Tübingen, Center for Applied Geoscience, WESS, Hölderlinstr. 12, 72076 Tübingen, Germany

ARTICLE INFO

Article history:
Received 26 May 2015
Received in revised form 31 August 2015
Accepted 25 December 2015
Available online 5 January 2016
This manuscript was handled by Peter K. Kitanidis, Editor-in-Chief, with the assistance of Roseanna M. Neupauer, Associate Editor

Keywords:
Statistics
Land use
Groundwater quality
Categorical variable
Mixed distribution
Inverse problem

SUMMARY

Most groundwater recharge comes from the infiltration of water through the land surface. Data analysis shows that solute concentrations at the water table vary between land use categories and depending on the land use composition within a certain neighbourhood.

Driven by these observations, the goal of this paper is to estimate the solute distribution at a location depending on the composition of land use in the neighbourhood, even though land use information is categorical. This goal is achieved by mixing pure distributions of homogeneous land use according to their frequency of occurrence in the vicinity of, and their distance from an estimation location. These pure distributions are jointly inverted using a maximum likelihood-based approach.

The neighbourhood size is optimized using cross-validation. Measurements below detection limit are included via their probabilities of non-exceedance. A solute-specific, spatially distributed measure of information content of the secondary information is presented. The method is applicable for many types of secondary information and can be used as drift for spatial estimation of the primary variable. This estimation is a local estimation and does not include larger scale spatial information. The information of measurements is included via the optimized concentration distributions for land use groups, not via a model of spatial dependence. The global estimation is described in the companion paper.

1. Introduction

Secondary information can be used to improve the estimate of the primary variable. The objective of this paper is to develop a methodology to estimate a primary variable at unsampled locations using a secondary variable that (1) is not only a point measurement but rather an optimized mixture based on the composition of that secondary variable within a certain neighbourhood and (2) that is of categorical type.

The primary variables of this study are groundwater quality parameters measured at the water table and the secondary variable are land use categories, as land use effects groundwater quality. Despite the obvious necessity, the effects of land use on groundwater quality are analysed relatively sparsely. Foley et al. (2005) pointed out that the effects of land use on groundwater quality exist: they presented a detailed assessment of the possible effects of land use change on food production, freshwater and forest resources, regional climate and air quality, and infectious diseases. Meiyappan and Jain (2012) estimate that in the last 250 years more than half of the global ice-free land has been modified by humans. Hydrogeologically, it is clear that land use must have an impact on groundwater quality and recharge rates, as most groundwater originates from excess rainfall infiltrating through the land surface (Foster and Cherlet, 2014). These authors also point out that groundwater response to land use impacts will usually be gradual and often delayed due to the large storage capacity of most aquifer systems. Lerner and Harris (2009) go as far as assessing the extent and effects of planning tools such as source protection zones on land use. Scanlon et al. (2005) worked on quantifying the impact of land use change on groundwater recharge and quality by detailed analysis of vertical water movement within the unsaturated zone, distinguishing between irrigated and dry land agriculture. Some studies exist that quantify the effects of land use more generally in hydrology, for example on flooding (O'Connell et al., 2007; Bronstert et al., 2002). Groundwater analysis is typically performed using numerical models at the scale of wellhead protection areas (Haslauer et al., 2005).

* Corresponding author. Tel.: +49 (0)7071 29 73081.
E-mail address: claus.haslauer@uni-tuebingen.de (C.P. Haslauer).

http://dx.doi.org/10.1016/j.jhydrol.2015.12.049
0022-1694/

Nomenclature

k	$k = 1, \dots, K$ observations, censored and not censored	n	$n = 1, \dots, N$ location in the vicinity of target
i	$i = 1, \dots, I$ observations, not censored	l_k	land use at location x_k
j	$j = 1, \dots, J$ observations, censored	d	Euclidean distance
g	$g = 1, \dots, G$ land use group	α_n	weight at location n
s	$s = 1, \dots, S$ supporting point	h	kernel density
ϕ_{sg}	weight of group g at supporting point s	H	kernel distribution
γ_{kg}	weight of group g in the vicinity of point k	v_g	frequency of occurrence of group g
x	coordinates of a location k	U	uniformly distributed variable
z	observed value/interpolation quantity	b	constant
0	target/interpolation location		

The methodology demonstrated in this paper utilizes land use as secondary information to enhance the estimation of anthropogenic groundwater quality at unsampled locations. For example, one would expect that nitrate concentrations in groundwater under farm fields are generally larger than under forests due to the excessive application of fertilizers. Groundwater quality parameters are influenced by anthropogenic, biogenic, and geogenic processes. For shallow groundwater systems, anthropogenic processes are dominant, hence the composition of land use is expected to provide useful information for the estimation of groundwater quality parameters close to the land surface. This paper presents methodologies to incorporate information about categorical land use data within the vicinity of an interpolation location to estimate the contaminant distribution at any location where measurements may not necessarily exist. There are two important factors to consider: (1) not only a mean, but a full distribution is estimated at any location, and (2) not only will the land use directly at the interpolation location be used, but the composition of the land use within the vicinity of every interpolation location. This distribution is referred to in this paper as "locally mixed distribution", whereas the distributions of the land use groups are referred to as "global" or "pure" distribution functions. Ultimately, the locally mixed distributions will lead to an improved estimation of spatially distributed parameters relevant for hydrology.

There are two key problems that are solved with the presented methodologies: (1) Land use is categorical information. Based on categories alone it is not possible to derive or estimate the distribution of the primary variable at a location where it was not measured; (2) The composition of the land use in the vicinity of an interpolation location should be considered. That means that the value of the secondary information directly at the interpolation location alone is not sufficient for a good estimate, but instead a neighbourhood with a certain size and associated land use composition should be evaluated. These two basic hypotheses have four main implications: (a) The distribution of the secondary information at an unsampled location must be a mixture of the pure distributions of the categories that exist within a given neighbourhood. The distribution of the secondary information at an unsampled location is then allowed to vary at each location as the composition of the neighbourhood varies; (b) the pure distributions for a given neighbourhood size are unknown and must be estimated (this is an inverse problem); (c) the number of categories of secondary information should be minimized. This number becomes more and more important as larger neighbourhoods are considered: the larger the number of categories, the tougher the inverse problem is to solve; (d) The size and the shape of the neighbourhoods are subject to optimization.

Our hypothesis was that the distribution of a primary variable can be estimated at an unsampled location by a mixed distribution reflecting the composition of the secondary information within the neighbourhood of a certain size and shape. The distributions used for mixing and the ideal neighbourhood size are subject to optimization. For anthropogenic contaminants, these locally mixed distributions represent the contribution of vertical infiltration. This information could be used in a copula-based geostatistical framework together with the horizontal transport component to interpolate groundwater quality.

Intuitively, a certain area with its individual composition of land uses should effect the concentration of a contaminant at a continuous shallow groundwater table at a given point. This composition is different in different parts of the domain, hence the result is a "local" distribution. Because the composition of the land use in the vicinity of a given point is incorporated, it is also called "mixed". The resulting distribution that is influenced by the composition of the secondary information in its vicinity is hence being referred to as "mixed local distribution". For example, the nitrate concentration at an interpolation location that lies within a forest, but is close to a farm field, is expected to be influenced by potentially larger nitrate concentrations under the farm field, and also by the potentially smaller nitrate concentrations under the forest. Such an approach based on mixture distribution has the benefit of smoother and more realistic boundaries. The size and the shape of the neighbourhood are subject to optimization. The underlying distributions used for mixing are "pure" distributions, i.e. distributions of a homogeneous neighbourhood composed of one group of secondary information. These distributions vary with neighbourhood size and are generally unknown and need to be estimated via inversion. Rarely, a sufficient number of measurements exists with large enough uniform neighbourhoods to estimate these distributions.

In this study, the focus is not on censored measurements (e.g. measurements below detection limit), but they are included in the estimation of the marginal distribution. It has been long accepted that censored measurements contain useful information that should be used in (low-dimensional statistical) models – see amongst others Cohen (1976), Helsel and Gilliom (1986), Helsel and Cohn (1988), Liu et al. (1997), Kroll and Stedinger (1999) and Shumway and Azari (2002) for reference.

This paper is structured in the following way: First it is demonstrated based on the data used, that there is considerable variability of contaminant concentrations within land use classes and that differently composed neighbourhoods lead to different contaminant concentrations. Following these motivational statements, the methodology of employing non-parametric distribution functions to describe measurements is presented and extended to incorporate censored measurements (Section 3). Secondly, the process of merging similar distribution functions is shown (Section 4), which is needed to reduce the parameter space. The concept of mixed distributions is explained in Section 5, which is needed when the "pure" distribution functions (pure in the sense of a single land use group within a given neighbourhood size) need to be optimized (Section 6). Finally, tests for determining an ideal