付宇, 曹文庚, 张娟娟. 基于随机森林建模预测河套盆地高砷地下水风险分布[J]. 岩矿测试,2021,40(6):860 – 870.
FU Yu, CAO Wen – geng, ZHANG Juan – juan. High Arsenic Risk Distribution Prediction of Groundwater in the Hetao Basin by Random Forest Modeling[J]. Rock and Mineral Analysis,2021,40(6):860 – 870.
【DOI: 10.15898/j.cnki.11 – 2131/td.202108170099】

基于随机森林建模预测河套盆地高砷地下水风险分布

付宇[1], 曹文庚[2*], 张娟娟[3]
(1. 华北水利水电大学, 河南 郑州 450046;
2. 中国地质科学院水文地质环境地质研究所, 河北 石家庄 050061;
3. 河北省地矿局第六地质大队, 河北 石家庄 050085)

摘要: 河套盆地浅层地下水砷污染严重,对当地居民健康造成严重影响。当前对河套盆地浅层地下水高砷分布的研究受限于采样时间和样本数量,难以从宏观角度对河套盆地高砷地下水的空间分布作出较为全面的评价。本文基于研究区 506 个浅层地下水样品,以 9 个地表环境参数为初始预测变量,经过最佳变量组合筛选,采用随机森林建模来产生风险概率,评价了预测变量的重要性以及对高砷地下水的影响。以气候因子为动态预测变量,根据模型识别不同季节地下水高砷的概率分布并制作了风险区专题图。结果表明:研究区的地下水样品砷含量为 0.05 ~916.7μg/L,超标率(砷浓度 >10μg/L)为 50%;地下水高砷风险区主要分布在河套盆地的沉积中心地带,但冬季高砷风险区面积减少 1907km^2,占研究区总面积 14.14%;降水、干旱指数、排灌渠影响、潜在蒸散、温度是影响高砷地下水最重要的指标。研究认为,河套盆地的气候变量(降水、干旱指数)与含水层砷含量显著相关,控制高砷地下水在河套盆地的沉积中心地带发生季节性变化。

关键词: 地下水; 砷污染; 河套盆地; 随机森林; 季节变化; 风险分布

要点:

(1) 建立随机森林模型,以气候因子为动态驱动,识别不同季节高砷概率分布。

(2) 气候变量(降水量、干旱指数)与含水层中的砷积累显著相关。

(3) 地下水砷高风险区集中于河套盆地的沉积中心地带,冬季高砷风险区面积小于夏季。

中图分类号: P641 **文献标识码:** A

地下水在全世界的灌溉用水和饮用水供应中发挥着重要作用,然而全球范围内已有七十多个国家检测出高砷地下水。世界卫生组织(WHO)发布的饮用水中砷的临时指导值为 10μg/L,人们长期饮用高砷水(>10μg/L)会引起各种皮肤病、癌症和心血管疾病[1]。同时,高砷灌溉水也会导致作物中产生高无机砷从而引发人类疾病[2-3]。高砷地下水最早于 20 世纪初在阿根廷被报道,随后在印度、孟加拉、柬埔寨、中国、越南、缅甸、美国发现了分布面积更广、砷浓度更高的高砷区,目前全球高砷地下水影响人口高达 1.4 亿[4-5]。在中国,1500 万人的健康受到高砷地下水的威胁,特别是新疆、内蒙古、山西、宁夏等干旱 – 半干旱地区[6]。研究高砷地下水的分布及驱动因素,对当地居民饮水安全以及地下水资源的合理利用具有重要现实意义。

全球有许多主要含水层存在严重的地下水砷污染问题。尽管这些地区各自具有特定的砷来源、水文过程、地质沉积构造等条件,但高砷地下水的发

收稿日期: 2021 – 08 – 17; **修回日期:** 2021 – 09 – 08; **接受日期:** 2021 – 09 – 21
基金项目: 国家自然科学基金项目(41972262); 河北自然科学基金优秀青年科学基金项目(D2020504032); 河南省高校重点科研项目计划 (19A170010)
第一作者: 付宇,博士,讲师,从事地质信息化工作。E – mail: fuyu1203@163.com。
通信作者: 曹文庚,博士,副研究员,从事水文地球化学、水文地质工作。E – mail: 281084632@qq.com。

生主要分布在两类环境中:①降水丰富、补给量大的三角洲平原。如以孟加拉为代表的南亚地区;②干旱 - 半干旱的内陆沉积盆地。如中国河套盆地、银川盆地、大同盆地等。河套盆地作为中国境内典型的富砷内陆沉积盆地,地下水砷含量严重超标,最高达到 1480μg/L[7],超过 30 万人的身体健康受到威胁,饮水型地方性砷中毒患病率达 15.54%[8]。为了查明河套盆地高砷地下水的成因、分布、富集、迁移机制,国内外众多学者在该区域开展了调查研究[9-12],已获得极为丰富的成果。研究表明河套盆地以还原环境为主的沉积环境、特定的地质条件、构造环境是地下水砷异常的原因[7,12-14]。近十年来,郭华明研究团队[6,11-12,15-16]从微观层面揭示了河套盆地砷的富集和迁移是伴随着硫酸盐、铁氧化物、氢氧化物的还原而发生,同时这一过程还受到含水层土著微生物、天然胶体、参与反应的有机物以及地下水开采作用的影响。目前对微观层面的河套盆地水文生物地球化学过程有了较完整的认识,但未见从宏观角度对河套盆地高砷地下水的空间分布与驱动机制作出较为全面的评价。而从宏观角度对大尺度空间范围的高砷地下水分布进行预测,可以帮助识别区域地下水中可能含有高浓度砷的地区。研究不同时段高砷地下水分布演变,对河套盆地高砷地下水动态演化机制研究具有一定的参考价值。前期已有学者实现了大尺度空间范围地下水砷的预测。如美国地质调查局于 2006 年率先报道了对美国整个东北部地区基岩含水层高砷(>5μg/L)地下水使用逻辑回归模型进行风险预测[17]。Rodríguez - Lado 等[18]利用逻辑回归模型预测了中国华北地区及江汉平原可能存在高砷地下水,并判断出地形、土壤为主要控制因素。Wu 等[19]使用随机森林模型绘制了印度地下水中砷浓度(>10μg/L)的分布图和高砷暴露人口分布图,并发现了尚待确定的潜在高砷区。

预测地下水污染物分布的机器学习模型包括:逻辑回归模型[20]、神经网络模型[21]、支持向量机模型[22]、随机森林模型[23]等。随机森林算法具有调参少、预测精度高和泛化能力强等优点,且不易产生“过拟合”现象,对异常值和噪声具有很好的容忍度,对特征选取具有较好的鲁棒性,是数据分类和预测的普遍选择[24]。随机森林模型能发现目标变量和预测变量之间的统计关系,以便进行预测[20,23]。这种方法可以用来考虑各种与地下水中砷的释放和积累有关的环境因素,指出模型中环境因素和目标变量之间的关系。针对砷在沉积含水层中的高度不均匀分布,可基于二元目标变量建模来产生高砷概率分布,为后续对整个区域高砷地下水成因解读提供科学、可靠的依据。

为了动态识别不同季节地下水中可能含有高浓度砷的地区,分析地下水砷的变化,明确重要影响因子的驱动作用,本文基于河套盆地 506 个浅层地下水样品,以气候变量为动态驱动,利用随机森林建模方法对河套盆地高砷地下水空间分布进行建模,识别冬夏季高砷区分布,进而评价各预测变量在砷浓度预测过程中的重要性,进一步研究气候变量对高砷地下水分布预测的影响。

1　研究区概况

河套盆地位于内蒙古自治区的西部,其北部为阴山山脉,南邻黄河,西接乌兰布和沙漠,东侧是淡水湖乌梁素海。行政区属巴彦淖尔市,面积 1.2 万 km^2(图 1)。地势平坦开阔,局部起伏,形成岗丘和洼地。该盆地属于温带大陆性干旱 - 半干旱季风气候,光照丰富,昼夜温差较大,降水量的季节分配不均,降水量远低于蒸发量且 70% 左右的降水集中在夏季,年平均降水量 130 ~ 220mm,年平均蒸发量 1900 ~ 2500mm[25]。

河套盆地是一个中新生代断陷盆地,在地质构造上属于华北地台鄂尔多斯台向斜的北缘部分。河套盆地的第四系全新统和上更新统($Q_4 + Q_3$)通常为冲洪积、冲湖积沉积物,岩性主要为夹砾粗砂、细砂、粉砂等,含水层厚度大(可达 200m 以上),分布广,因此具有较大的供水意义。高砷地下水正是分布在这一含水层中,对当地居民的健康造成了较大的影响。

盆地地形总体趋势是自东南向西北降低,地面海拔为 1070 ~ 1030m。地下水位在 1.5 ~ 20m 左右,地下水流速从山前地区到平原区逐渐递减[26]。盆地内浅层地下水水位埋深,除了阴山山前冲洪积扇裙带水位埋深大于 5m,盆地大部分地区地下水位埋深在 5m 之内[7],地下水的多年动态变化规律主要受到黄河水灌溉的影响,地下水运移以垂向入渗为主,侧向径流微弱。地下水的补给来源主要有降水、地表水、灌溉水的垂直入渗及山前侧向径流等方式。由于盆地气候类型为干旱 - 半干旱内陆季风气候,自然降水少且蒸发强烈,对地下水的补给效果较差,降水入渗占总补给量的 23.1%。盆地内部广泛分布着不同级别的引黄灌溉渠网,引黄灌溉水的入渗补给是地下水主要的补给形式,灌溉水占总补给量

的 76.5%[27]。盆地内部天然形成的地表水体较少,盆地内的排干主要为地下水的排泄路径。

2　实验部分

2.1　样品采集与分析测试

河套盆地采样数据来自 2016 年 9 月进行的水文地质调查,采样井在研究区均匀分布,位置如图 1 所示。调查采集了浅层地下水样品 506 组(井深 2~120m,样品采集深度以采样井滤水管中间位置为准),样品采集深度基本都控制在晚更新世含水层。

采样前先测量地下水位,清洗井孔,抽出大于井孔储水量 3 倍以上的水量(对于常用且日用水量较大的井直接抽水取样,无洗井),待 pH 值等现场测试指标稳定后进行样品采集。现场测定的指标如水温、pH、电导率、溶解氧、氧化还原电位,使用美国哈希 sension2 台式离子浓度计和 Quanta 便携式水质测定仪测定;地下水中易被氧化的 NH_4^+、NO_3^-、S^{2-} 和 Fe^{2+} 的含量,使用美国哈希 DR2800 便携式分光光度计测得。用于砷元素分析的地下水水样采集均使用 0.22μm 滤膜现场过滤,过滤的上清液装入 25mL 高密度聚乙烯棕色采样瓶,采样瓶预先用 10% 硝酸浸泡 12h,并使用去离子水清洗 6 次。样品需滴加 1mL 浓盐酸,将水样酸化至 pH<2。用于分析砷形态的水样装于 2mL 棕色玻璃瓶中,并加入 0.25mol/L 乙二胺四乙酸保存。所有采样瓶密封保存在 0~4℃冷藏箱中,7 日之内送回实验室进行测试。

样品测试工作由中国地质科学院水文地质环境地质研究所承担。检测环境温度 23℃,湿度 50%。砷元素采用美国 Aglient 公司 7500C 电感耦合等离子体质谱仪(ICP-MS)测试;地下水样品中砷的形态分析以美国 PerkinElmer 公司 200B/785A/TURBO EL HPLC SYSTEM 型液相色谱仪、Pecosphere C18 色谱柱以及北京瑞利公司 AF-610 原子荧光光谱仪为硬件平台,通过高效液相色谱-氢化物发生-原子荧光光谱法(HPLC-HG-AFS)测定。分析地下水样时,加 5% 的平行样品,所有平行样品的误差小于 5%,表明各项指标的准确性均在质量要求范围内。

2.2　模型建立

2.2.1　变量选择

建模的目标是距离地表一定深度的地下水砷浓度,在获取环境变量数据过程中,由于大范围内获取地下空间数据(如地球物理测量、钻井)的成本普遍较高,难度较大,只有地表空间数据获取方式简单(遥感影像、地表监测站),而且在时间和空间上是连续的。在建模过程中,时间和空间连续的数据更能反映出预测目标的时空变异性特征。

本次建模考虑了“气候”、“地形”和“其他”三类,共计 9 个地表空间连续数据用作建模的预测变量(表 1)。这些变量的选择是基于其在地下水砷的累积过程中已知或潜在的功能。在干旱-半干旱地区,强烈的蒸发浓缩过程是影响该地区水化学特征的主要水文地球化学过程[25],该区高砷地下水的形

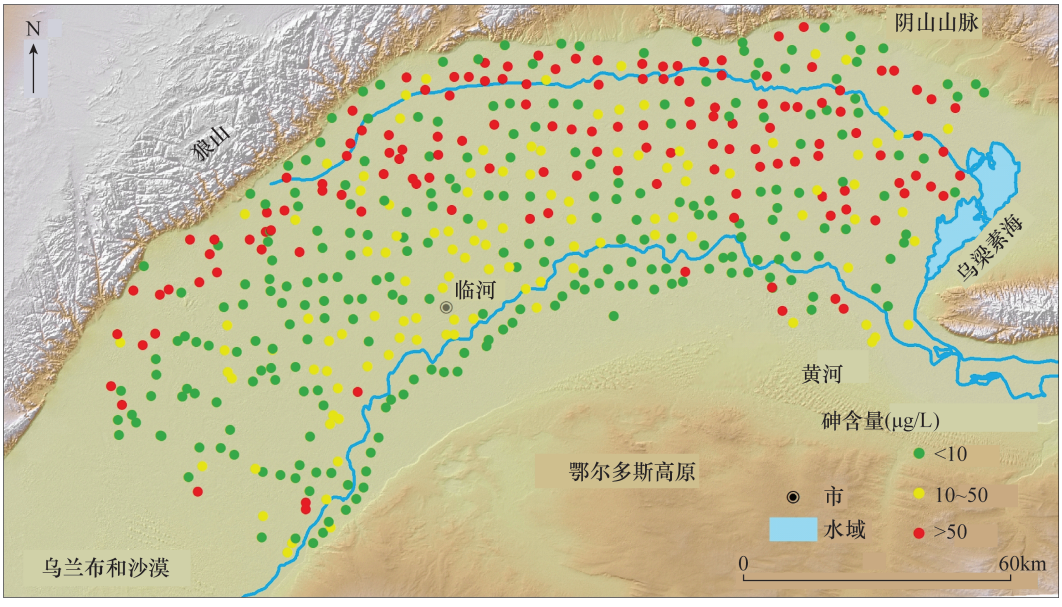


图 1　河套盆地地下水砷含量分布

Fig. 1　Distribution of arsenic content in groundwater in Hetao Basin

成属于自然环境水文地球化学作用为主的成因类型[8]。高存荣[14]研究表明河套地区高砷地下水的形成与特定的地质、地形条件有关。研究区广泛存在的灌渠以及排干影响了地下水的流动,对砷的富集也产生影响,在灌渠及排干的附近,地下水更偏向于形成还原环境,有利于砷的富集;远离这些地表水体的位置,还原条件增强,水砷含量升高[28]。

为了剔除表现不佳的预测变量,创建最佳模型,通过交叉验证的方式执行递归特征消除(RFE)迭代生成初始 9 个特征变量的子集,使用随机森林算法来计算所有子集的验证误差,选择误差率最小的特征子集,将该特征子集中的变量作为最终的建模变量。9 个变量中具有时间连续性的预测变量(蒸散、降水等)选用夏季(6、7、8 月)数据,最终模型选取的是 9 个变量的集合,这是错误率最小且在随机森林一个标准误差范围内的模型。

表 1　模型预测变量及描述

Table 1　Predictor variables and descriptions of the model

类别	变量	描述
气候	真实蒸散	平均真实蒸散量(mm/mr)
	干旱指数	温度植被干旱指数
	潜在蒸散	平均潜在蒸散量(mm/mr)
	降水	平均降雨量(mm/mr)
	温度	平均温度(℃/mr)
地形	高程	单位为 m
	坡度	单位为(°)
其他	排灌渠影响	排干、灌渠影响
	植被指数	归一化植被指数

2.2.2　建模与验证

利用上述砷浓度数据集和预测变量,建立地下水中砷含量超 10μg/L 的统计预测模型。

首先对比了一些统计学习方法,包括 Logistic 回归、支持向量机与随机森林方法。选择均方根误差(RMSE)、平均绝对误差(MAE)和平均相对误差(MRE)三个评价指标进行分析,RMSE、MAE、MRE 值越小,表明模型越优,预测精度越高。随机森林方法在对比中表现出最好的预测性能。

随机森林模型是 Breiman[29] 提出的一种基于 CART (Classification and Regression tree)决策树的组合模型,主要有分类(RFC)和回归(RFR)两种算法,基本思想是基于统计学理论提出的,通过自助(Boot - strap)重采样技术,从原始训练样本集 N 中有放回地重复随机抽取 K 个样本生成新的训练样本集合,然后根据自助样本集生成 K 个决策树组成的随机森林。此外,每棵树可用的预测变量是随机选择的,而且数量受到限制。由于并非所有变量都同时考虑,在随机森林模型中,通常可以忽略预测因子之间多重共线性问题。对于分类模型,新数据的分类结果按分类树投票的多少而定,而对于回归模型,将所有决策树预测平均值作为最终预测结果。

模型中需设置的参数主要有:决策树数目;树节点划分时随机选取的预测变量数目。理论上,决策树数目越大,预测精度越高;树节点划分时随机选取的预测变量数目是模型最敏感的参数,通常是取预测变量的平方根。通过均方误差与决策树数目的关系、袋外数据(Out of Bag,OOB)误差与树节点划分时随机选取的预测变量数的关系,来确定最终决策树数目和树节点划分时随机选取的预测变量数。结果表明,当决策树数目超过 400 时,均方误差基本趋于稳定状态;当树节点划分时随机选取的预测变量数为 4 时,OOB 误差达到最小值。因此,选取决策树数目为 400,决策树节点划分时随机选取的预测变量数为 4 作为最优参数。

由于地下水砷在沉积含水层中的分布往往高度不均匀,因此通常采用基于二元目标变量的建模来产生概率[19],用 10μg/L 作为阈值。首先根据砷浓度 <10μg/L 或 >10μg/L 或等于 10μg/L,将砷浓度重新编码为 0 或 1。利用砷浓度编码集和上述自动选择过程确定的预测变量(夏季),将 506 个采样点数据集随机分为训练(80%)和测试(20%)数据集。使用训练数据集建立模型,然后将该模型应用于测试数据集,通过各种统计数据评估其性能,以确定其在预测新数据的低($\leqslant$10μg/L)和高($>$10μg/L)砷浓度方面的准确性。

性能评估参数包括受试者工作特征(ROC)曲线下面积(AUC)以及准确率(Accuracy)。AUC 表征预测高值(灵敏度)和低值(特异性)的准确性,它是通过对建模的概率在 0 和 1 之间应用许多不同的阈值得出的,通过计算特异性与灵敏度绘制的曲线下的面积得到 AUC 值,该值通常在 0.5(未经证实的猜测)到 1(完美的预测精度)之间。准确率是针对所有的测试数据而言的,表征有多少样本被准确预测,它是通过计算测试数据集中预测正确的正类(TP)和负类(TN)在所有预测数据中的占比得出,公式表达为:Accuracy = (TP + TN)/(TP + TN + FP + FN),式中 FP 和 FN 分别代表假正类和假负类。

最后,将该模型应用于这 9 个空间上连续的预测变量集,创建河套盆地不同季节地下水中高砷浓度概率图作为风险区的分布图。

3　结果与讨论

3.1　地下水砷含量统计和分布特征

对研究区内地下水砷含量描述性统计特征分析,所有地下水采样点中砷含量为 0.05 ~ 916.7 μg/L,平均值为 58.35μg/L,中位值为 9.43μg/L,以世界卫生组织(WHO)发布的饮用水中砷的临时指导值 10μg/L 为标准,该区域地下水 506 个样点砷含量的超标率为 50%,具有很高的人体健康风险。所有样点含量的变异系数为 1.97,表明研究区地下水砷含量具有很强的空间变异性。从砷含量的数据范围也可以看出,砷含量极差较大,数据集不符合正态分布特征,存在右偏尾现象。

从区域上(图 1)来看,高砷区呈带状分布,且分布规律与前人研究成果一致[7-8]。地下水砷含量超过 50μg/L 的区域自西向东呈条带状,并逐渐向南扩展,西部的高砷地下水沿狼山山前冲洪积扇缘的低洼地带呈北东向的条带状分布;东部区以五原为中心,高砷地下水区多呈不规则的片状分布,范围较广,砷含量最大值可达 916.7μg/L。

3.2　地下水高砷风险分布特征

模型在测试数据集上的交叉验证结果如表 2 和表 3 所示。该随机森林模型在测试数据集上的整体准确率为 0.7426,显著高于无信息率 0.5545($p = 7.338 \times 10^{-5}$)。无信息率是指在没有预测模型的情况下所能达到的精度,即数据集中数据比例较大的类别所占的比例,56% 的砷测量点等于或大于 10μg/L。同样,Kappa 统计量(0.4767)是一个超出偶然预期的精度指标,通常 Kappa 值在 0(不一致)与 1(完全一致)之间变化。ROC 曲线下的面积(AUC)为 0.784,通常 AUC 值的范围为 0.5(没有预测能力)到 1(完美的预测能力),同时 AUC 还可以代表在众多概率截断值中,二元模型预测低值和高值的能力[30]。

在 ROC 曲线中根据各点对应的敏感性和特异性,计算(敏感性 + 特异性 - 1)获取最大值的点作为概率截断点 0.509,可用于确定地下水砷浓度的高风险区域。与其他国家或地区的地下水砷预测模型对比,本研究区(河套地区)建立的模型 AUC 值和准确度是一个比较理想的结果。例如,印度古吉拉特邦的 AUC 值为 0.71 ~ 0.83[19],印度北方邦的 AUC 值为 0.74[31],巴基斯坦的 AUC 值为0.8[32],美国中北部的准确率为 0.67[33],中国山西省的准确率为 0.68[34]。

表 2　随机森林模型的混淆矩阵(概率截断值 = 0.5)

Table 2　Confusion matrix of the random forest model (probability cutoff = 0.5)

预测类别	真实类别	
	0	1
0	31	12
1	14	44

表 3　随机森林模型的统计数据(概率截断值 = 0.5)

Table 3　Statistics data of the random forest model (probability cutoff = 0.5)

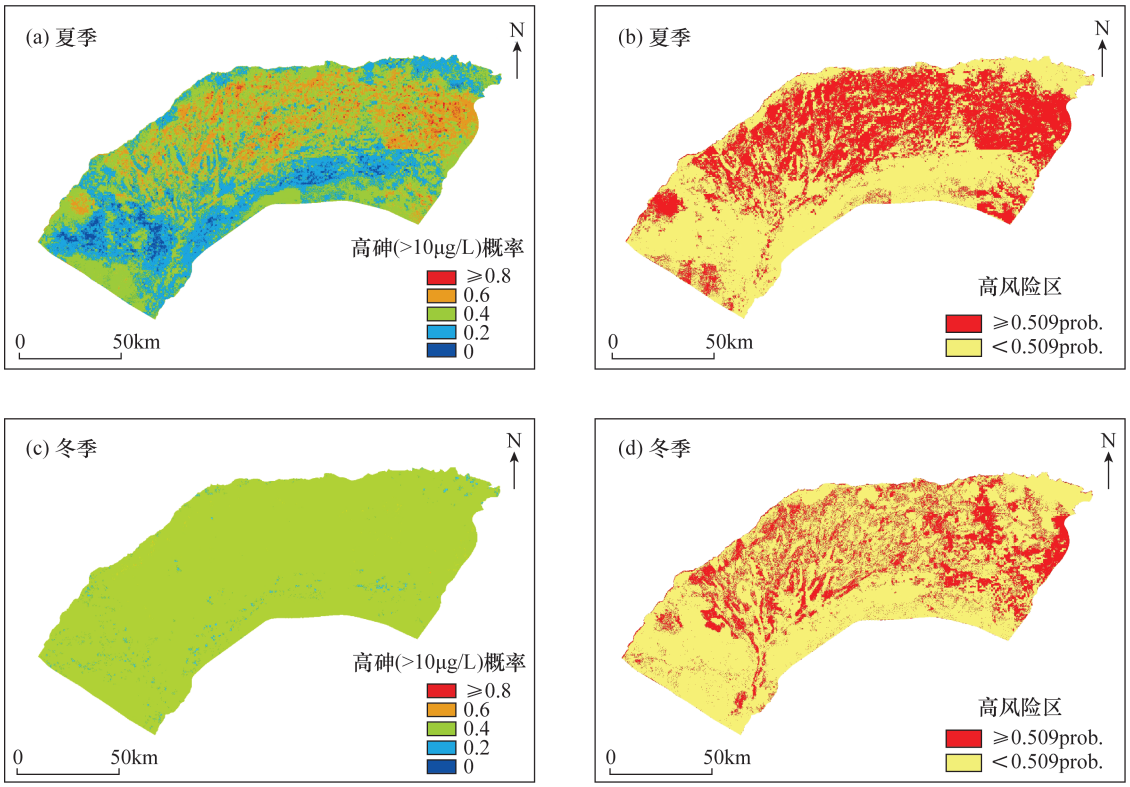
评估参数	参数数值	评估参数	参数数值
准确率	0.7426	特异性	0.7857
无信息率	0.5545	阳性预测值	0.7209
p	7.338×10^{-5}	阴性预测值	0.7586
Kappa 系数	0.4767	流行度	0.4455
敏感性	0.6889	平衡精度	0.7373

最后,使用随机森林模型分别计算了河套盆地夏季和冬季砷浓度超过 10μg/L 的概率。地下水高砷概率图如图 2 中 a 和 c 所示,结合概率阈值0.509(>10μg/L),绘制了地下水砷高风险分布图如图 2 中 b 和 d 所示。

夏季高砷地下水高风险区,涵盖了黄河古河道影响带、黄河北岸决口扇的冲积沉积物中已知的高砷地区,同时还涵盖了部分没有获取砷浓度数据的地区,根据概率阈值划定的高砷面积达到 5571km^2,占研究区范围的 38.73%。高砷概率范围在 0.04 ~ 0.91之间,其中概率大于 0.6 的区域则更为集中分布在河套盆地的沉积中心地带。

冬季高砷地下水高风险区,根据概率阈值划定的高砷面积达到 3665km^2,占研究区范围的 24.59%,主要分布于排干沿线及乌梁素海西侧低洼地带。相比夏季,冬季高砷面积减少了 1907km^2,减少面积占全区总面积的 14.14%。高砷概率范围在 0.28 ~ 0.65 之间,其中概率大于 0.6 的区域则零星地分布在排干沿线及乌梁素海西侧低洼地带。

河套盆地降雨量少蒸发量大,通常情况下 70% 左右的降水集中在夏季,本次研究采用的 2016 年降水数据,夏季降水基本囊括了全年降水量,约为



(a)夏季和(c)冬季地下水中砷浓度超过 10μg/L 的概率; (b)夏季和(d)冬季基于概率截断点 0.509 的高危险区。

图 2　不同季节砷风险分布

Fig. 2　Arsenic hazard maps by season. Probability of arsenic concentration in groundwater exceeding 10μg/L in (a) summer and (c) winter. High hazard areas based on probability cutoffs of 0.509 in (b) summer and (d) winter

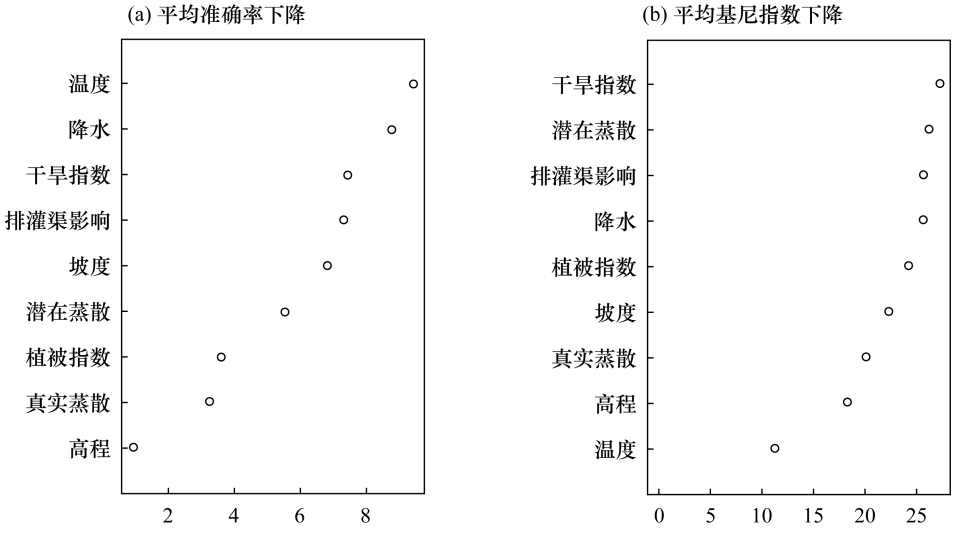
120mm 左右。夏季集中降水致使地下水水位上升,使空气无法进入地层而形成还原环境,在干旱的气候环境条件下地表水的 pH 值普遍偏高,这样给地层中砷的溶出提供了有利的条件。冬夏两季由于气候条件的变化可能导致冬夏高砷区空间分布差异。已有研究表明,地下水 As 浓度的季节性变化也有类似的结果。在汉江平原,雨季(6 ~ 9 月)砷浓度逐渐升高,雨季结束时(9 月)砷浓度达到最高,随后砷浓度开始逐渐下降,最低值是在旱季结束时(4 ~ 5 月)[35]。如 Yadav(2015)[36]对印度恒河流域上游的观测值显示,砷浓度的时空分布与季节相关,冬季较低,夏季较高。

3.3　高砷概率与预测变量关系分析

预测变量的重要性被用来帮助评估不同预测变量对模型的相对影响。随机森林模型中预测变量的重要性评估主要使用两种统计方法:①精确度降低;②基尼节杂质减少。在最终的模型中,不同预测变量对研究区地下水的砷风险概率影响的重要性程度及排序如图 3 所示。所有的预测因子都没有负值,这表明它们都对模型有益。在精确度和基尼系数的平均下降方面,对其每一项都进行了标准化,综合这两种统计方法的结果,得出降水、干旱指数、排灌渠影响、潜在蒸散、温度、植被指数是对模型预测影响重要性排序靠前的预测变量,其中降水和干旱指数对地下水砷含量空间分布模拟的准确性贡献度最大。

根据预测变量重要性综合排序可得出,最终模型中的气候变量(降水量、干旱指数、潜在蒸散量和温度)与含水层中的砷积累显著相关,表明气候对含水层砷释放的强大控制作用。高温促进了蒸发蒸腾,加剧了干旱。高蒸散量、高干旱指数、高温和低降雨量的结合会增加地下水的蒸发量,从而增加砷的浓度,特别是在干旱或半干旱气候下的内陆或封闭盆地[37]。河套盆地作为一个较封闭的内陆盆地,其氧化还原电位(Eh)显示出几乎为负值的还原环境,在脱硫酸作用下,使 pH 值升高,胶体吸附力减弱,游离砷含量增加[8]。降水和蒸散发在这种还原以及高 pH 值的干旱环境下,有利于砷释放条件的产生[38]。

离排干的距离也影响着地下水砷的富集。一项针对总排干附近地下水的研究[28]发现,高砷含量的



a—将 9 个变量的取值变为随机数时,模型预测准确性的降低程度; b—9 个变量对分类树每个节点上观测值的异质性的影响程度。该值越大,表示该变量的重要性越大。

图 3　随机森林模型预测变量重要性排序

Fig. 3　Importance ranking of predictor variables in the random forest model. (a) Represents the reduction of model prediction accuracy when the values of 9 variables are changed into random numbers; (b) Represents the influence degree of nine variables on the heterogeneity of observed values at each node of the classification tree. The greater the value, the greater the importance of the variable

底泥的物源保证、灌溉输入磷肥带来的竞争吸附以及灌溉季水位抬升导致的还原环境,可造成总排干附近地下水砷富集。大量黄河水的引入不仅使地下水水位抬升,形成大面积的土壤盐渍化,而且地下水水位的上升使空气无法进入地层形成还原环境,给地层中砷的溶出提供了有利的条件。引黄灌溉使地下水位和地下水压力发生显著的变化,这样进入土粒空隙中的砷很容易进入地下水中。

已有研究表明,蒸散量会随植被覆盖度的增加逐步增加,因此地表植被覆盖度对地下水砷含量也可能存在影响[39-40]。区域地形平缓,代表着地下水径流极其缓慢,有助于抑制地下水系统对砷的冲刷,同时还促进了较细粒的冲积沉积物、含砷铁氧化物和沉积物中丰富有机质的积累。然后,在含水层微生物的作用下将砷释放到地下水中,导致地下水的砷赋存于地下水流动缓慢的平坦、低洼地区[15,19,26]。

4　结论

通过利用随机森林建模,识别河套盆地不同季节高砷地下水的潜在风险,分析了预测变量对地下水砷浓度的影响,并分析重要指标对高砷地下水的作用。结果表明:①浅层地下水砷含量的超标率(砷浓度 > 10μg/L)为 50%,高砷区集中地分布在河套盆地的沉积中心地带;②气候因子(降水、干旱指数)在对预测模型重要性评估中占主导地位,以气候因子为动态驱动,对不同季节的高砷地下水开展风险预测切实可行;③冬季地下水砷高风险区面积比夏季减少 1907km^2,减少的面积占全区总面积的14.14%,主要分布于排干沿线及乌梁素海西侧低洼地带。

本研究建立的气候驱动下的高砷地下水地表空间参数模型,达到了初步预期效果,对于大范围内获取地下空间参数难度较大的情况,本研究成果可提供一定的参考价值。但是,河套盆地高砷地下水分布也会受到不同季节灌溉条件变化的影响,本模型中并未涉及灌溉情况。模型的预测质量取决于它们所基于的数据,由于地下水砷局部尺度具有显著的空间变异性,为了获得最稳健的高砷地下水砷分布结果,还需要更详细的预测变量(沉积环境、土壤、水化学、灌溉)和更多的砷浓度数据集的加入,对该模型实现进一步改进。

5　参考文献

[1] Polya D A, Middleton D R. Arsenic in drinking water: Sources & human exposure [M]//Bhattacharya P, Polya D A, Draganovic D. Best practice guide on the control of arsenic in drinking water (The 1st edition). London:

International Water Association Publishing,2017.

[2] Tong J,Guo H,Wei C. Arsenic contamination of the soil – wheat system irrigated with high arsenic groundwater in the Hetao Basin,Inner Mongolia,China[J]. Science of the Total Environment,2014,496:479 – 487.

[3] 杨文蕾,沈亚婷. 水稻对砷吸收的机理及控制砷吸收的农艺途径研究进展[J]. 岩矿测试,2020,39(4):475 – 492.
Yang W L,Shen Y T. A review of research progress on the absorption mechanism of arsenic and agronomic pathways to control arsenic absorption[J]. Rock and Mineral Analysis,2020,39(4):475 – 492.

[4] Wang Y,Pi K,Fendorf S,et al. Sedimentogenesis and hydrobiogeochemistry of high arsenic Late Pleistocene—Holocene aquifer systems[J]. Earth – Science Reviews,2019,189:79 – 98.

[5] Podgorski J,Berg M. Global threat of arsenic in ground – water[J]. Science,2020,368:845 – 850.

[6] Jia Y,Guo H,Jiang Y,et al. Hydrogeochemical zonation and its implication for arsenic mobilization in deep groundwaters near alluvial fans in the Hetao Basin,Inner Mongolia[J]. Journal of Hydrology,2014,518(Part C):410 – 420.

[7] 曹文庚,董秋瑶,谭俊,等. 河套盆地晚更新世以来黄河改道对高砷地下水分布的控制机制[J]. 南水北调与水利科技,2021,19(1):140 – 150.
Cao W G,Dong Q Y,Tan J,et al. The mechanism of Yellow River diversion in controlling high arsenic groundwater distribution since the Late Pleistocene[J]. South – to – North Water Transfers and Water Science & Technology,2021,19(1):140 – 150.

[8] 高存荣,刘文波,冯翠娥,等. 干旱半干旱地区高砷地下水形成机理研究:以中国内蒙古河套平原为例[J]. 地学前缘,2014,21(4):13 – 29.
Gao C R,Liu W B,Feng C E,et al. Study on the formation mechanism of high arsenic groundwater in arid and semi – arid areas: Taking Hetao Plain in Inner Mongolia as an example[J]. Earth Science Frontier,2014,21(4):13 – 29.

[9] Cao W G,Guo H M,Zhang Y L,et al. Controls of paleochannels on groundwater arsenic distribution in shallow aquifers of alluvial plain in the Hetao Basin,China[J]. Science of the Total Environment,2018,613 – 614:958 – 968.

[10] Guo H M,Li X M,Xiu W,et al. Controls of organic matter bioreactivity on arsenic mobility in shallow aquifers of the Hetao Basin,P. R. China[J]. Journal of Hydrology,2019,571:448 – 459.

[11] 李媛. 内蒙古河套盆地高砷含水系统的微生物特征及生物地球化学效应[D]. 北京:中国地质大学(北京),2016.
Li Y. Microbial characteristics and biogeochemical effects of high arsenic aquifer system in Hetao Basin,Inner Mongolia[D]. Beijing: China University of Geosciences (Beijing),2016.

[12] Shen M M,Guo H M,Jia Y F,et al. Partitioning and reactivity of iron oxide minerals in aquifer sediments hosting high arsenic groundwater from the Hetao Basin,P. R. China[J]. Applied Geochemtry,2018,89:190 – 201.

[13] Dietrich S,Bea S A,Weinzettel P,et al. Occurrence and distribution of arsenic in the sediments of a carbonate – rich unsaturated zone[J]. Environmental Earth Sciences,2016,75(2):1 – 14.

[14] 高存荣. 河套平原地下水砷污染机理的探讨[J]. 中国地质灾害与防治学报,1999(2):25 – 32.
Gao C R. Research on the mechanism of arsenic pollution in groundwater in the Hetao Plain,Inner Mongolia,China[J]. The Chinese Journal of Geological Hazard and Control,1999(2):25 – 32.

[15] Guo H M,Li Y,Zhao K,et al. Removal of arsenite from water by synthetic siderite: Behaviors and mechanisms[J]. Journal of Hazardous Materials,2011,186(2 – 3):1847 – 1854.

[16] Zhang Z,Guo H,Zhao W,et al. Influences of groundwater extraction on flow dynamics and arsenic levels in the western Hetao Basin,Inner Mongolia,China[J]. Hydrogeology Journal,2018,26(5):1499 – 1512.

[17] Ayotte J D,Nolan B T,Nucklos J R,et al. Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment[J]. Environmental Science & Technology,2006,40:3578 – 3585.

[18] Rodríguez – Lado L,Sun G,Berg M,et al. Groundwater arsenic contamination throughout China[J]. Science,2013,341:866 – 868.

[19] Wu R H,Joel P,Michael B,et al. Geostatistical model of the spatial distribution of arsenic in groundwater in Gujarat State,India[J]. Environmental Geochemistry and Health,2021,43:2649 – 2664.

[20] Bretzler A,Lalanne F,Nikiema J,et al. Groundwater arsenic contamination in Burkina Faso,West Africa: Predicting and verifying regions at risk[J]. Science of the Total Environment,2017,584 – 585:958 – 970.

[21] 苏彩红,向娜,陈广义,等. 基于人工蜂群算法 BP 神经网络的水质评价模型[J]. 环境工程学报,2012,6(2):699 – 704.
Su C H,Xiang N,Chen G Y,et al. Water quality

evaluation model based on artificial bee colony algorithm and BP neural network[J]. Journal of Environmental Engineering,2012,6(2):699 – 704.

[22] He Z B,Wen X H,Liu H,et al. A comparative study of artificial neural networks,adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semi – arid mountain region[J]. Journal of Hydrology,2014,509:379 – 386.

[23] Podgorski J E,Labhasetwar P,Saha D,et al. Prediction modeling and mapping of groundwater fluoride contamination throughout India[J]. Environmental Science and Technology,2018,52(17):9889 – 9898.

[24] Iverson L R,Prasad A M,Matthews S N,et al. Estimating potential habitat for 134 eastern US tree species under six climate scenarios[J]. Forest Ecology and Management,2008,254:390 – 406.

[25] 邓娅敏. 河套盆地西部高砷地下水系统中的地球化学过程研究[D]. 武汉:中国地质大学(武汉),2008.
Deng Y M. Study on geochemical process in high arsenic groundwater system in western Hetao Basin[D]. Wuhan:China University of Geosciences (Wuhan),2008.

[26] 袁溶潇. 内蒙古河套盆地含水层沉积物可溶性组分与可溶性砷的分布规律研究[D]. 北京:中国地质大学(北京),2017.
Yuan R X. Soluble components of sediments and their relation with soluble arsenic in aquifers from the Hetao Basin,Inner Mongolia[D]. Beijing:China University of Geosciences (Beijing),2017.

[27] 刘文波. 河套平原地下水化学特征研究[D]. 北京:中国地质大学(北京),2015.
Liu W B. Chemical characteristics of groundwater in Hetao Plain[D]. Beijing:China University of Geosciences (Beijing),2015.

[28] 何薪. 河套平原农业灌溉影响下地下水中砷迁移富集规律研究[D]. 武汉:中国地质大学(武汉),2010.
He X. Study on the migration and enrichment law of arsenic in groundwater under the influence of agricultural irrigation in Hetao Plain[D]. Wuhan:China University of Geosciences (Wuhan),2010.

[29] Breiman L. Random forests[J]. Machine Learning,2001,45(1):5 – 32.

[30] Fawcett T. An introduction to ROC analysis[J]. Pattern Recognition Letters,2006,27(8):861 – 874.

[31] Bindal S,Singh C K. Predicting groundwater arsenic contamination:Regions at risk in highest populated state of India[J]. Water Research,2019,159:65 – 76.

[32] Podgorski J E,Eqani S A M A S,Khanam T,et al. Extensive arsenic contamination in high – pH unconfined aquifers in the Indus Valley[J]. Science Advances,2017,3:1 – 10.

[33] Erickson M L,Elliott S M,Christenson,et al. Predicting geogenic arsenic in drinking water wells in glacial aquifers,North – Central USA:Accounting for depth – dependent features[J]. Water Resources Research,2018,54(12):172 – 187,10.

[34] Zhang Q,Rodríguez – Lado L,Johnson C A,et al. Predicting the risk of arsenic contaminated groundwater in Shanxi Province,northern China[J]. Environmental Pollution,2012,165:118 – 123.

[35] Michael V S,Samantha C Y,Shawn G B,et al. Aquifer arsenic cycling induced by seasonal hydrologic changes within the Yangtze River Basin[J]. Environmental Science & Technology,2016,50(7):3521 – 3529.

[36] Yadav I C,Devi N L,Singh S. Spatial and temporal variation in arsenic in the groundwater of upstream of Ganges River Basin,Nepal[J]. Environmental Earth Science,2015,73:1265 – 1279.

[37] Alarcón – Herrera M T,Bundschuh J,Nath B,et al. Co – occurrence of arsenic and fluoride in groundwater of semi – arid regions in Latin America:Genesis,mobility and remediation[J]. Journal of Hazardous Materials,2013,262:960 – 969.

[38] Joel P,Wu R H,Biswajit C,et al. Groundwater arsenic distribution in India by machine learning geospatial modeling[J]. Environmental Research and Public Health,2020,17:7119 – 7135.

[39] 张巧凤,刘桂香,于红博,等. 基于 MOD16A2 的锡林郭勒草原近 14 年的蒸散发时空动态[J]. 草地学报,2016,24(2):286 – 293.
Zhang Q F,Liu G X,Yu H B,et al. Temporal and spatial dynamics of evapotranspiration in Xilingole grassland in recent 14 years based on MOD16A2[J]. Journal of Grassland,2016,24(2):286 – 293.

[40] 闫俊杰,吕光辉,徐海量,等. 2000—2014 年塔里木河干流的植被覆盖与蒸散发时空变化及其关系[J]. 水土保持通报,2018,38(3):248 – 255.
Yan J J,Lv G H,Xu H L,et al. Temporal and spatial changes of vegetation cover and evapotranspiration in the main stream of Tarim River from 2000 to 2014 and their relationship[J]. Bulletin of Water and Soil Conservation,2018,38(3):248 – 255.

High Arsenic Risk Distribution Prediction of Groundwater in the Hetao Basin by Random Forest Modeling

FU Yu[1], *CAO Wen-geng*[2*], *ZHANG Juan-juan*[3]

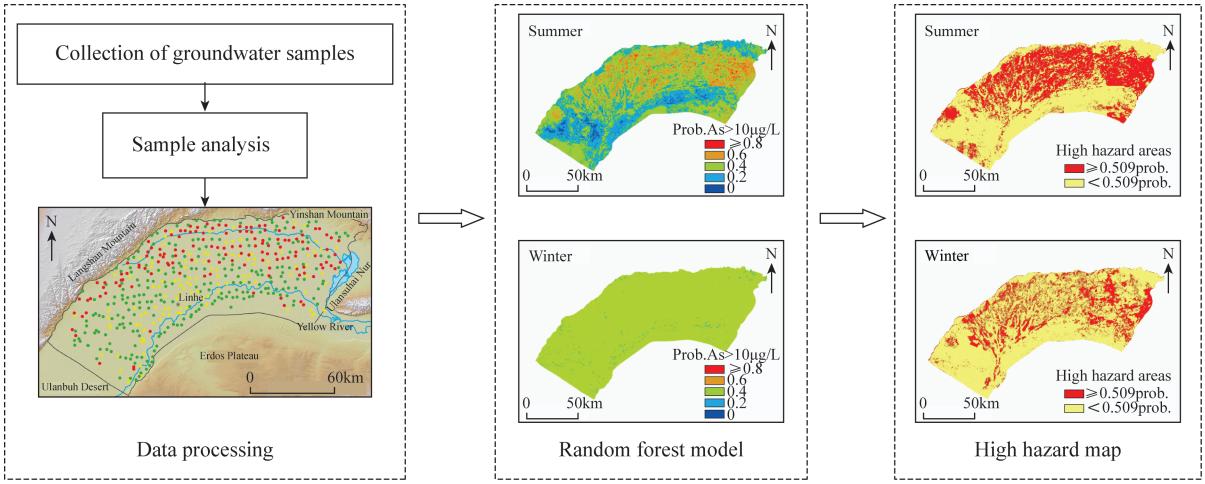
(1. North China University of Water Resources and Electric Power, Zhengzhou 450046, China;
2. Institute of Hydrogeology and Environmental Geology, Chinese Academy of Geological Sciences, Shijiazhuang 050061, China;
3. The 6th Geological Team, Hebei Bureau of Geology and Mineral Resources, Shijiazhuang 050085, China)

HIGHLIGHTS

(1) A random forest model was established to identify the probability distribution of high arsenic areas in different seasons driven by climate factors.

(2) Climate variables (precipitation and drought index) were significantly correlated with arsenic accumulation in aquifers.

(3) High arsenic risk areas in groundwater were concentrated in the depositional center of the Hetao Basin, and the area of high arsenic risk areas in winter was smaller than that in summer.



ABSTRACT

BACKGROUND: Arsenic pollution is a serious problem in shallow groundwater in the Hetao Basin, and has seriously affected the health of residents. The research on the distribution of high arsenic shallow groundwater in the Hetao Basin is limited by the sampling time and sample number.

OBJECTIVES: To obtain a comprehensive understanding of the risk distribution characteristics and important influencing factors of high arsenic groundwater in different seasons in the region.

METHODS: Based on 506 shallow groundwater samples and 9 surface environmental parameters as prediction variables, a random forest model was established to evaluate the importance of prediction variables and the impact of important variables on high arsenic groundwater. Taking the climate factors as the dynamic prediction variables, the probability distribution of high arsenic groundwater in different seasons was identified and thematic maps of risk areas were made.

RESULTS: The results showed that the arsenic content of 506 groundwater samples ranged from 0.05 to 916.7 μg/L with an overshoot rate (>10μg/L) of 50%. Groundwater arsenic risk areas were mainly distributed in the depositional center of the Hetao Basin, but the area of groundwater arsenic risk areas decreased by 1907km^2 in winter, accounting for 14.14% of the total area. Precipitation and drought index, influence of drainage and irrigation channels, potential evapotranspiration and temperature were the most important indexes affecting the high arsenic groundwater in this area.

CONCLUSIONS: In the Hetao Basin, climate variables (precipitation and drought index) are significantly correlated with arsenic accumulation in the aquifer, which controls the seasonal variation of groundwater with high arsenic content in the depositional center of the Hetao Basin.

KEY WORDS: groundwater; arsenic pollution; Hetao Basin; random forest; seasonal variation; risk distribution