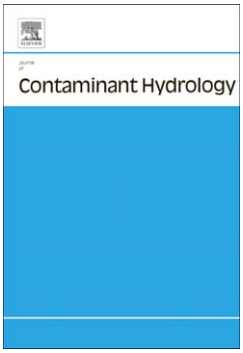

Groundwater source contamination mechanisms: Physicochemical profile clustering, risk factor analysis and multivariate modelling

Paul Hynds, Bruce D. Misstear, Laurence W. Gill, Heather M. Murphy

PII: S0169-7722(14)00025-4
DOI: doi: 10.1016/j.jconhyd.2014.02.001
Reference: CONHYD 2978

To appear in: *Journal of Contaminant Hydrology*

Received date: 18 August 2013
Revised date: 30 January 2014
Accepted date: 5 February 2014

Please cite this article as: Hynds, Paul, Misstear, Bruce D., Gill, Laurence W., Murphy, Heather M., Groundwater source contamination mechanisms: Physicochemical profile clustering, risk factor analysis and multivariate modelling, *Journal of Contaminant Hydrology* (2014), doi: 10.1016/j.jconhyd.2014.02.001

Groundwater source contamination mechanisms: Physicochemical profile clustering, risk factor analysis and multivariate modelling

Paul Hynds[1*], Bruce D. Misstear[1], Laurence W. Gill[1], Heather M. Murphy[2]

[1] Environmental Engineering Research Group, School of Engineering, Trinity College, Dublin 2, Ireland

[2] Formerly Centre for Foodborne, Environmental and Zoonotic Infectious Diseases, Public Health Agency of Canada, 255 Woodlawn Rd. West, Unit 120, Guelph, ON N1H 8J1

* Author to whom correspondence should be addressed; Email: hyndsp@tcd.ie

Abstract

An integrated domestic well sampling and "susceptibility assessment" programme was undertaken in the Republic of Ireland from April 2008 to November 2010. Overall, 211 domestic wells were sampled, assessed and collated with local climate data. Based upon groundwater physicochemical profile, three clusters have been identified and characterized by source type (borehole or hand-dug well) and local geological setting. Statistical analysis indicates that cluster membership is significantly associated with the prevalence of bacteria ($p = 0.001$), with mean *E. coli* presence within clusters ranging from 15.4% (Cluster-1) to 47.6% (Cluster-3). Bivariate risk factor analysis shows that on-site septic tank presence was the only risk factor significantly associated ($p < 0.05$) with bacterial presence within all clusters. Point agriculture adjacency was significantly associated with both borehole-related clusters. Well design criteria were associated with hand-dug wells and boreholes in areas characterized by high permeability subsoils, while local geological setting was significant for hand-dug wells and boreholes in areas dominated by low/moderate permeability subsoils. Multivariate susceptibility models were developed for all clusters, with predictive accuracies of 84% (Cluster-1) to 91% (Cluster-2) achieved. Septic tank setback was a common variable within all multivariate models, while agricultural sources were also significant, albeit to a lesser degree. Furthermore, well liner clearance was a significant factor in all models, indicating that direct surface ingress is a significant well contamination mechanism. Identification and elucidation of cluster-specific contamination mechanisms may be used to develop improved overall risk management and wellhead protection strategies, while also informing future remediation and maintenance efforts.

Keywords: Cluster Analysis, *E. coli*, Groundwater, Physicochemical Profile, Risk Factor Analysis

1. Introduction

Private groundwater sources in the Republic of Ireland are often poorly maintained, largely un-monitored and un-regulated; they are the primary drinking water supply for an estimated 210,000 households or 14-15% of the Irish population (CSO, 2012). Two principal private supply types exist in Ireland: private, unregulated groundwater supplies typically serving individual households; and private group water schemes serving <50 people (EPA, 2010). Similarly, large consumer numbers rely on private domestic wells in other developed countries; for example, McDonald *et al*. (2005) estimate that there are 20,000-30,000 private groundwater wells currently in use in Scotland, while an estimated 3-4 million Canadians (13%) are served by private supplies (Corkal *et al*., 2004; Charrois, 2010). Private domestic wells constitute the largest share of water wells in the United States — more than 13.2 million year-round occupied households have their own well, supplying 45 million people (US Bureau of the Census, 2008). The majority of aquifers in the Republic of Ireland are of composed of consolidated bedrock formations in which groundwater storage and transmission primarily occurs in fractures (i.e. secondary porosity). Accordingly, bedrock type is paramount with regard to the overall groundwater yielding potential of Irish aquifers (IGI, 2007). It is considered that limited attenuation of contaminants typically occurs in the bedrock due to the relatively rapid nature of fissure flow; thus, the overlying subsoils act as the primary protective layer for attenuating contaminants. Consequently, subsoil type and thickness are the most important natural features influencing groundwater vulnerability and groundwater contamination in Ireland (Swartz *et al*., 2003; Misstear & Fitzsimons, 2007). The two most important and common subsoil types in Ireland are glacial deposits (tills) and glaciofluvial sand and gravel deposits.

A large proportion of private groundwater sources, both in Ireland and abroad, are situated in rural areas; accordingly, a multitude of potential point and non-point pathogen sources are present. These include diffuse agricultural sources such as grazing animals and land-spreading, point agricultural sources including farmyards, silage pits and animal housing and point sources such as on-site wastewater treatment systems, solid waste landfill sites and road runoff (Borchardt *et al.,* 2003; Giannoulis *et al.*, 2005; Rozemeijer & Broers, 2007). There are two main processes by which domestic wells may become contaminated: generalized aquifer contamination and localized "source-specific" contamination (Godfrey *et al*., 2006). Hynds *et al*. (2012) further classify localized mechanisms as being related to direct ingress at the wellhead or due to rapid and/or shallow groundwater pathways. Hence, areas dominated by low permeability subsoils may exhibit higher than expected levels of domestic well contamination, particularly in the case of poorly designed or constructed wells, due to direct ingress of contaminants at the wellhead. Localised pathways may be developed through poor design, construction and/or operation of private groundwater supplies,

particularly areas with low permeability soils and subsoils that generate high runoff. More generalized groundwater pathways may exist due to the hydrogeological setting in a particular area, including bedrock type, subsoil type and depth (groundwater vulnerability) and aquifer importance (Godfrey *et al.,* 2006; Hynds *et al.*, 2012). Aquifers can be classified in terms of their importance using multiple criteria (Payne & Woessner, 2010; Olaniyan *et al*., 2010). In Ireland, aquifer classification is based primarily on the overall aquifer productivity (potential) yield and areal extent of the aquifer. These factors are amalgamated to formulate an overall aquifer value as a groundwater resource (Table 1). There are three main aquifer categories (Regionally important, Locally important and Poor aquifers) defined in Groundwater Protection Schemes (DoELG/EPA/GSI, 1999), further subdivided into nine categories (Table 1).

Based upon previous work by the authors (Hynds *et al.*, 2012), the objective of the current study is to further elucidate domestic well contamination mechanisms in diverse geological regions, using the Republic of Ireland as a case-study. To date, little research has focused specifically on this topic. The overall study approach and findings will aid water managers, well users and local government in improving risk management decisions and developing evidence-based quantitative wellhead protection strategies. Moreover, where contamination has occurred, results may be used to inform remediation and maintenance efforts.

2. Materials and Methods:

2.1 Study Areas

In all, 211 private groundwater sources over four study areas were assessed, sampled and included in the current study (Figure 1). Study areas were selected using developed inclusion/exclusion criteria to maximise (hydro)geological and source type representivity with respect to the Republic of Ireland. Descriptions of the selection procedures and the study areas have been given previously in Hynds *et al.* (2012). Inclusion/exclusion criteria included: groundwater vulnerability, availability of previous monitoring data, well density, laboratory proximity and hydrogeological mapping status. Three areas of *High* or *Extreme* groundwater vulnerability were selected for study. Additionally, one area categorized as *Low* vulnerability was selected for comparative purposes (Table 2).

2.2 Sample Analysis

Groundwater samples were obtained in accordance with Standard Methods (APHA/AWWA/WEF, 2005), while all on-site analyses were undertaken in accordance with USGS National Field Manual for the Collection of Water Quality Data (USGS, 2005). Sterilised 500 ml glass bottles were used for sample collection, whereupon samples for *E. coli* analysis were immediately

transferred to a cooler and transported to a laboratory. Time between sample collection and microbial analysis did not exceed 6 h. All physicochemical parameters were measured in the field and recorded. Groundwater temperature was taken with a conventional laboratory thermometer (^{o}C). A WTW ProfiLine 197i™ Portable Conductivity Meter was used for all electrical conductivity (EC) analyses (µS cm^{-1} at 25 ^{o}C). A Cyber-Scan Series 600™ Waterproof Portable Meter was used for all pH analyses. EC and pH meter calibration was undertaken at the start of each day in the field at reference EC values of 700 and 2000 µS cm^{-1}, and pH values 4 and 7.

Isolation and enumeration of *E. coli* were performed using the membrane filtration (MF) culture method in accordance with standards methods (APHA/AWWA/WEF, 2005).Bacterial growth took place on sterile membrane filters (white, grid marked, 47 mm diameter, 0.45 µm) saturated with membrane lauryl sulphate broth (MLSB; buffered at pH 7.4 ± 0.2). Plates were incubated for 4 h at 25^{o}C to aid resuscitation, before transferring to 44^{o}C ± 0.5^{o}C for 16 h (Oxfam-DelAgua, 2004). All yellow colonies with a diameter of 1-3 mm were counted as *E. coli* and recorded. Confirmation tests carried out during this study (ONPG-MUG preparations and indole production in UV light) suggest that >90% of coliform colonies isolated were *E. coli* spp.

2.3 Risk Factor Data

As many site- (i.e. site on which well is located) and well-specific details as possible were recorded upon sample collection in order to collate all potentially relevant risk factors with respect to groundwater contamination. Data were recorded under three primary assessment types: hydrogeological characterization, site characterization and source characterization (Hynds *et al*. 2012). A detailed list of collated potential risk factors, employed metrics and methods is presented in Appendix 1.

2.4 Precipitation Data

Antecedent total daily precipitation data were obtained from the synoptic station closest to the respective study area via Met Eireann (The Irish Meteorological Service). Precipitation data (mm) were summed for the 24 h (1 day), 48 h (2 days), 120 h (5 days) and 30 day (monthly) antecedent periods preceding all sampling events for risk factor and multivariate analysis.

2.5 Statistical Analyses

All statistical analyses were performed within the R statistical environment (R Development Core Team; i386 2.15.2); significance was determined using a p-value <0.05 (Agresti, 1996). Associations between potential risk factors and confirmed *E. coli* presence were analyzed to inform working hypotheses and subsequent multivariate analyses. Results of these analyses were used to explore and correct for (multi)-collinearity among potential risk factors. Due to differing risk factor

variable types, a number of bivariate techniques were used as appropriate. In the case of *E. coli* presence/absence (dichotomous variable), chi-square tests of independence, Mann-Whitney U tests and independent samples t-tests were employed when analysing nominal, ordinal and continuous risk factor datasets, respectively. One-way analysis of variance (ANOVA) was used to determine and quantify associations between continuous and categorical variables, with Bonferroni post-hoc multiple comparison analysis employed upon rejection of the global null hypothesis H_0 (Agresti, 1996).

2.5.1 Cluster Analysis

Analyses were carried out to determine whether definable subgroups existed with respect to groundwater physicochemistry. Agglomerative hierarchical clustering (between groups linkage; squared Euclidean distance method) was used to optimize suitable cluster numbers based upon continuous physicochemical variables and sample size. Hierarchical cluster analysis indicated that three clusters appropriately characterised the dataset. Schwartz Bayesian Information Criterion (SBIC) clustering was employed, using a standard two-step algorithm. SBIC clustering was used as the number of input parameters was low and because it has been widely applied for model identification, particularly within regression analyses (Lindberg *et al.,* 2010). Clustering diagnostics included: cluster centroid analysis, simultaneous cluster mean variation and cluster-wise parameter importance using Bonferroni adjustment.

2.5.2 Multivariate Analysis

Binary Hierarchical Logistic Regression (HLR) was employed to predict the occurrence of *E. coli* in private wells in Ireland. This approach was taken because the prediction of specific *E. coli* concentrations (cfu/100ml) is not realistic due to unquantifiable variability within the system (e.g. magnitude of *E. coli* in faeces of particular animal species, high resolution temperature variation affecting microbial die-off, specific on-site geological flowpaths). MANOVA and multivariate regression analysis were not considered appropriate as the primary response variable (*E. coli*) included numerous “non-detects” (i.e. <1 cfu/100ml) and a discrete threshold value was specified to define the response categories.

The logistic equation examines how the probability of an event changes as the predictor variables change, via estimation of parameters. This allows for optimisation of the model by obtaining maximum likelihood estimates of parameters using an iterative-reweighted least squares algorithm (Hosmer & Lemeshow, 2000; Helsel & Hirsch, 2002). Using this iterative process, values are calculated to maximize the log-likelihood function L (*Eqn. 1*):

$$L = \sum_{i-1}^{m}[(y_i \times lnP_i) + (1 - y_i) \times \ln(1 - P_i) \qquad \textit{Eqn.1}$$

where,

m = number of observations in the dataset

y_i = outcome variable set to 1 when *E. coli* ≥1 cfu/100ml; otherwise a value of 0 is assigned

P_i = *probability of contamination*

Potential risk factors of importance were entered into logistic regression (LR) models hierarchically, based upon perceived importance from risk factor evaluation. A stepwise method of model entry (Forward: Conditional) was used for predictor variables, with variable entry taking place at p = 0.05 and removal at p = 0.1. A maximum of 20 estimation iterations were run for each model step, with estimation terminated where parameter estimates changed by less than 0.001 between two estimation iterations. After model development, significant predictor parameters in the model were assigned to associated identifiable hierarchies. Models were subsequently re-run with these hierarchies, so that the relative importance (i.e. proportion of system variability and overall significance of the hierarchy) could be quantified. A classification cut-off of 0.5 was used as this was the most conservative approach (i.e. classification cut-off < 0.5 = *E. coli* absent; *E. coli* > 0.5 = *E. coli* present). A similar analytical approach has been employed previously by Hynds *et al*. (2012).

3. Results

3.1 Physicochemical Clustering

Three physicochemical parameters (groundwater temperature, pH, and EC) were used as cluster inputs as these effectively account for source depth, surface water interactions, seasonality and geological setting (OCM, 2007). Both EC and pH are referred to as lithologically influenced "non-global" parameters, thus, both have characteristic ranges for differing local geological settings (OCM, 2007). As shown (Table 3), Cluster-1 was characterised by low temperature, low pH and low EC, Cluster-2 was described by comparatively high temperature, neutral pH and high EC, while Cluster-3 was observed as being relatively equidistant between Clusters 1 and 2 for all three variables. Analyses indicate that cluster membership and source type were significantly associated (χ^2 (2) = 208.0, p < 0.001) with Cluster-3 comprised entirely of hand-dug wells (n = 42). No significant association was found between Cluster-3 membership and any (hydro)geological input variables (e.g. bedrock type, subsoil type, subsoil thickness, aquifer importance or groundwater vulnerability category).

No significant associations were found between cluster membership and borehole design (i.e. wellhead finish, wellhead clearance, flange clearance, wellhead cover, site conditions or

chamber presence). However, significant associations existed between cluster membership and (hydro)geological variables, including vulnerability category (χ^2 (4) = 52.06, p < 0.001), bedrock type (χ^2 (13) = 140.9, p < 0.001), subsoil type (χ^2 (5) = 118.6, p < 0.001) and aquifer importance (χ^2 (4) = 53.33, p < 0.001). Sources in Cluster-1 were located in areas dominated by Dinantian limestone bedrocks with limestone tills and Devonian sandstone bedrocks with sandstone tills, while boreholes in Cluster-2 were typically dominated by granites and Lower Paleozoic bedrocks and associated subsoils. The majority (88%) of sources in Cluster-1 were located in areas where the underlying aquifer was classified as being locally important, while Cluster-2 was dominated by poor aquifers (62%).

A significant association was found between cluster membership and *E. coli* presence (χ^2 (2) = 14.895, p = 0.001), thus clusters not only represented source type and geology, but were also useful for source susceptibility classification within the collated dataset. As shown (Table 3), low-level *E. coli* presence was associated with Cluster-1, while the highest level of *E. coli* presence was noted among sources within Cluster-3. However, no statistical relationship was found between *E. coli* magnitude (where present) and cluster membership.

3.3.2 Risk Factor Analysis

3.2.1 Cluster-1 (Boreholes in areas dominated by high/moderate permeability subsoils)

As shown (Table 4), risk factor analysis indicates that the primary hazard sources within Cluster-1 were point contamination sources (septic tanks and point agricultural sources); both measured septic tank setback distance and the recorded presence of a point agriculture source within a 100m radius of the wellhead were significantly associated with increased bacterial presence. Well design parameters were also significantly associated with *E. coli* presence, particularly wellhead finish (i.e. above, under or level with ground surface) and liner clearance (measured distance from top of well liner to ground surface or concrete chamber floor). Under- and over-ground wellhead finish were significantly associated with a lower likelihood of *E. coli* presence than wells finished at ground level (i.e. 0 mm liner clearance). Increased 120-h antecedent precipitation was associated with an increased probability of *E. coli* presence.

3.2.2 Cluster 2 (Boreholes in areas dominated by low permeability subsoils)

Similar to findings from Cluster-1, the primary potential hazard sources of significance with regard to bacterial contamination in Cluster-2 were point sources, namely septic tank systems and point agricultural sources (Table 5). Hydrogeological variables in Cluster-2 were also significantly associated with *E. coli* presence including both bedrock and subsoil type (Table 5). A small number of

sources within Cluster-2 (n = 5) were underlain by limestone tills (although not associated with limestone bedrock), with 80% of these (n = 4) exhibiting bacterial presence; this figure was <12% (n = 9) among sources not underlain by limestone tills. Boreholes located in areas with <3 m subsoil cover were bacterially-contaminated in 20% of cases, while this was approximately 9% where subsoil thickness was >3 m. Neither antecedent rainfall volume nor well design criteria was statistically associated with *E. coli* presence.

3.2.3 Cluster 3 (Shallow hand-dug wells)

Bedrock type and aquifer importance were identified as being significant risk factors within Cluster-3 (Table 6); for example, 80% of hand-dug sources in sandstone areas were contaminated, compared with 28.5% in areas underlain by granites or other igneous intrusives. Further, a higher level of *E. coli* presence was observed in hand-dug wells located in locally important aquifers (65%) than poor aquifers (26%). Both road setback gradient and distance were significantly associated with bacterial presence; hand-dug wells with evidence of bacterial contamination had a mean road setback distance of 32.8 m, compared with a mean setback distance of 74 m among uncontaminated hand-dug wells. Septic tank systems were isolated as being the potential hazard source of greatest significance, with septic tank setback distance in particular identified as being significantly associated with contamination. Additionally, hand-dug sources supplying one household were observed as being more likely to have *E. coli* present than those supplying greater than one (OR 7.518, 95% CI 1.26 – 39.91).

3.3 Multivariate Analysis

3.3.1 Cluster-1

The developed HLR model included four significant input hierarchies;

1. Septic tank setback distance,
2. Well liner clearance (mm),
3. Agricultural point source vicinity (<100m),
4. Adjacent roadway setback gradient.

While 120-hr antecedent precipitation (mm) was found to improve the overall significance of the final Cluster-1 model (increased explanatory power of approximately 4%), it was not a significant parameter within the final developed model. The Hosmer-Lemeshow (H/L) diagnostic statistic for the final model (Table 7), was not significant (p = 0.104); thus, the final model is considered a good fit. A low Nagelkerke R^2 (0.463) indicates that the model explains approximately

46% of variability within the system, and could therefore be improved upon if more data were available (Table 10).

3.3.2 Cluster-2

In the case of Cluster-2, multivariate modeling resulted in the identification of four significant hierarchies (Table 8), as follows;

1. Septic tank setback gradient,
2. Subsoil thickness,
3. Agricultural grazing land setback gradient,
4. Wellhead design

Similar to the Cluster-1 model, the inclusion of 24-hour precipitation increased the model's overall predictive accuracy; however, explained variance was only increased by 1.7% to 0.614 and as coefficient significance within the model was >0.05, it was omitted. Similarly, grazing land setback distance and the presence of agricultural point sources within 100m of the wellhead increased the prediction accuracy but neither were significant in the final model. Both the -2 log-likelihood and H/L diagnostic indicate that the model is significant and provides a good fit for the input variables (Table 10).

3.3.3 Cluster 3

The final Cluster-3 multivariate model comprised four significant hierarchies (Table 9);

1. Septic tank setback distance and gradient,
2. Liner clearance (mm),
3. Agricultural point source vicinity (<100m),
4. Households supplied

When 24-hour precipitation was added to the multivariate model, the Nagelkerke R^2 increased to 0.851 and overall prediction accuracy was increased to 92.9%, with 95% and 90% of *E. coli* absence and *E. coli* presence correctly predicted, respectively. However, as this coefficient was not statistically significant within the overall model, it was omitted. Similarly, both grazing land setback distance and road setback distance explained additional system variance, but neither was significant within the model (Table 10).

4. Discussion

The authors have previously developed a "contamination susceptibility" model for private domestic wells in the Republic of Ireland, predicting the presence of thermotolerant coliforms in private wells with approximately 85% accuracy (Hynds *et al*., 2012). Subsequently, this approach has been further developed by employing statistical techniques to identify three clusters based on measured groundwater physicochemistry. These clusters characterize private wells within the same database according to source type (bored or hand-dug) and local hydrogeological setting. Further analysis indicates that the identified clusters can also distinguish source susceptibility with respect to confirmed *E. coli* presence. Consequently, both bivariate and multivariate analyses were applied to examine the likely dominant contamination mechanisms within each cluster. Statistical models that examine the interactions between groundwater contamination and relevant explanatory variables (e.g. geology, landuse, climate, topography, etc.) have been developed for a number of contaminants including pesticides (Teso *et al.*, 1996), nitrates (Nolan, 2001) and arsenic (Zhang *et al*., 2012). However, to date, few studies have used these methods to predict the possible presence of enteric pathogens. Models predicting contamination likelihood may be used to accelerate/aid "screening" of wells by prioritizing sampling (Zhang *et al*., 2012). Moreover, where contamination has occurred, these models may be used to indicate the likely route of contaminant ingress, thereby aiding source remediation/ingress (decrease future susceptibility). Additionally, they may be useful in informing future source construction i.e. differing wellhead design may be appropriate/necessary in differing hydrogeological settings (Zhang *et al*., 2012), particularly in regions where private groundwater supplies remain unregulated such as Ireland, Canada and numerous US states.

Based on characteristic pH and EC ranges for Irish aquifers (OCM, 2007), in concurrence with collated hydrogeological data and cluster centroid analyses, identified clusters were classified as: bored wells in areas underlain by high or moderate permeability subsoils with aquifers often comprised of limestones and sandstones (Cluster-1), bored wells underlain by areas dominated by low permeability subsoils and granitic aquifers (Cluster-2) and shallow hand-dug wells (Cluster-3).

The lack of significance between Cluster-3 membership and any collated geological variable (groundwater vulnerability category, aquifer importance, bedrock type, subsoil type, subsoil thickness) indicates Cluster-3 membership is solely based on well design and construction characteristics. Notably, while numerous studies have investigated the contamination of shallow hand-dug wells in developing countries, including Bangladesh (Hossain & Sivakumar, 2006; Luby *et al*., 2008), Uganda, (Howard *et al*., 2003; Kulabako *et al*., 2007) and Mozambique (Godfrey *et al*. 2006), few if any studies have examined shallow dug wells as a potential human source of enteric infection in more developed regions.

A mean overall *E. coli* presence of 29.4% was recorded over the 2-year sampling period within the four outlined study areas (Table 3). These figures are representative of Irish groundwater quality, for example, during the 3-year period 2007-2009, 30.3% of EPA groundwater monitoring locations (64/211) recorded faecal coliform presence, with 34.8% of all groundwater samples (945/2718) during the same period testing positive for faecal coliforms (EPA, 2010). A significant association was found between cluster membership and *E. coli* presence, with shallow hand-dug wells (Cluster-3) shown to exhibit an *E. coli* prevalence significantly greater than the mean prevalence rate. Conversely, bored wells in regions dominated by high or moderate permeability sediments displayed an *E. coli* prevalence rate well below the mean. This agrees with previous studies in which well depth has been highlighted as a significant risk factor associated with bacterial contamination (e.g. Tabbot & Robson, 2006; Gonzales, 2008).

Bivariate risk factor analyses were used to examine within-cluster bivariate relationships between *E. coli* presence and potential causative factors including hazard sources and geological or source-specific pathways (Tables 4-6). Findings indicate that point contaminant sources predominate, with septic tank setback (distance and/or gradient) significantly associated with *E. coli* presence within all clusters. Decreased septic tank setback distance was related to an increased likelihood of *E. coli* being present upon sampling within all clusters. For example, within Cluster-3, contaminated hand-dug wells had a mean setback distance of 62 m, whereas uncontaminated wells had a recorded mean setback of 134 m. This finding concurs with several previous studies (Beller *et al.*, 1997; Borchardt *et al*., 2003). Borchardt *et al.* (2010) reported that old and/or poorly maintained septic tank systems are likely causes of groundwater contamination, and consequently, large waterborne disease outbreaks. Fong *et al*. (2007) have previously reported on a large groundwater-related AGI outbreak affecting approximately 1,450 residents and visitors of South Bass Island, Ohio. A thorough epidemiological investigation found that septic tank systems were the primary enteric pathogen source. Approximately half of the world's population currently resides in rural areas and relies on a private on-site domestic wastewater treatment system, while 30% of European citizens utilize a septic tank for domestic wastewater treatment (WHO/UNICEF, 2010; Gunnarsdottir *et al*., 2013).

Previous work by Gaut (2005) on 49 private wells in Norwegian crystalline bedrock aquifers found significant correlations between microbiological water quality and (i) wellhead completion, (ii) type and thickness of superficial deposits, and (iii) landuse and contamination sources. Furthermore, Hrudey & Hrudey (2007) have reported that multiple mechanisms have been found to have contributed to waterborne outbreaks. Multivariate modelling and subsequent interpretation

presented here indicate that similar associations and "multiple-mechanism" contamination exist in private wells in Ireland.

Hierarchical logistic regression modelling within Cluster-1 indicates that the primary hazard sources associated with boreholes in limestone/sandstone areas dominated by high or moderate permeability subsoils are point sources, including both septic tank systems and point agricultural sources. These accounted for 17% and 37% of explained model variance, respectively, thus, point agricultural sources are shown to be the dominant hazard source within this cluster. The lack of significance of geological/hydrogeological parameters within the model was expected, due to inherent geological homogeneity within borehole-related clusters. Thus, the effects of site-specific geological and/or hydrogeological parameters remain "hidden", representing a limitation of the current approach. The lack of significance of short-term precipitation periods (24-hour, 48-hour) indicates that rapid ingress is not a dominant contamination mechanism among boreholes in these limestone/sandstone areas, likely due to moderately or highly permeable limestone/sandstone tills in these areas resulting in decreased runoff coefficients (Misstear *et al*., 2009). The significance of liner clearance, which accounted for 23% of explained model variability (Table 7), in concurrence with improved overall model accuracy associated with 120-hour precipitation (and its significance as a bivariate risk factor), implies that ingress due to shallow groundwater infiltration is the most likely contaminant transport mechanism within this cluster.

The developed hierarchical model indicates that bacterial contamination among boreholes in areas characterized by low permeability subsoils is associated with both point (septic tank systems – 32% variability) and non-point (grazing animals – 15.5% variability) sources (Cluster-2, Table 8); thus, both rapid recharge and direct wellhead ingress the likely mechanisms of contaminant transport, with direct ingress the dominant mechanism. Higher levels of bacterial presence in those sources associated with <3 m overlying subsoils suggests relatively rapid recharge through thin subsoil layers. The significance of wellhead finish, liner clearance and wellhead ground condition (34%) indicate direct ingress of contamination. The significance of short term precipitation (24-hour) also reflects a relatively rapid transport mechanism, particularly at poorly constructed or maintained wellheads. The significance of on-site conditions and wellhead radius conditions as risk factors in Cluster-2 indicates that "stewardship issues" may be associated with source susceptibility; therefore, consumer risk perception and awareness is an element of source susceptibility within Cluster-2. Future guidance on well maintenance should therefore include on-site and wellhead vicinity upkeep.

The model developed for hand-dug wells (Cluster-3, Table 9) was noted as being somewhat similar to the final model for Cluster-1 sources, with point contamination sources including both septic tank systems (42% variability) and agricultural point sources (14% variability), found to be the

primary hazard sources. The significance of liner clearance (21% variability) within the model suggests that direct wellhead ingress is a contamination mechanism of importance, with the association, albeit insignificant within the final model, of 24-hour precipitation also pointing to rapid direct ingress. An independent samples t-test found that hand-dug sources associated with >1 supplied household (n = 12; 28.5%) were less likely to have bacterial contamination present (OR 2.131). It is unclear why this is the case, however, this may be attributable to higher levels of maintenance associated with increased source shareholder numbers or increased rates of abstraction resulting in higher well throughput and therefore, decreased contaminant residence times (Schijven *et al*., 2006). No geological/hydrogeological factors were deemed to have significant explanatory value with regards bacterial contamination of hand-dug wells even though they were heterogeneous with respect to geological setting.

5. Conclusions

Multivariate analytical techniques were used to identify three groundwater source clusters based upon measured physicochemical profiles. Cluster membership was significantly associated with confirmed *E. coli* presence, thus indicating that clusters may be effectively employed as groundwater source susceptibility surrogates in the absence of sufficient data. *E. coli* prevalence was highest within the hand-dug well cluster; these source types and their inherent susceptibility have received a great deal of attention in developing countries. However, few if any studies have focused on their potential human health burden in more developed regions. Risk factor analyses and cluster-specific logistic models indicate that while similar hazard sources are significant within all clusters, with point sources in particular playing a central role, contamination mechanisms differ. Multivariate modelling suggests sources within Cluster-1 are characterised by localised mechanisms, namely relatively rapid direct ingress and more gradual shallow infiltration. Both generalised and localised mechanisms would seem to be causative within Cluster-2, with results suggesting that localised mechanisms are dominant, particularly rapid ingress after high intensity short duration rainfall events. Rapid ingress at the wellhead is likely to be the main mechanism of contamination among hand-dug wells (Cluster-3), with increased “flushing” and/or improved wellhead maintenance potential protective factors. It is considered that the simple statistical approach presented here may be employed to provide insights on well contamination mechanisms in diverse geological settings, including both developed and developing regions of the world. Thus, potential human health burdens attributed to contaminated groundwater sources may be prevented through improved source-specific management and remediation actions.

Acknowledgements

The authors gratefully acknowledge the support of the Environmental Protection Agency (EPA, Ireland) for funding this research under the Science, Technology, Research and Innovation for the Environment (STRIVE) 2007-2013 programme. The authors would also like to acknowledge the helpful comments provided by two anonymous reviewers.

References

Agresti A. (1996). Categorical Data Analysis. John Wiley & Sons, Inc. Hoboken, NJ

APHA/AWWA/WEF (2005). Standard Methods for the Examination of Water and Wastewater, (21st Ed). American Public Health Association, Washington, D.C.

Beller M., Ellis A., Lee SH., Drebot MA., Jenkerson SA., Funk E. (1997). Outbreak of viral gastroenteritis due to a contaminated well. International consequences. *JAMA.* **278**:563–568

Borchardt MA., Bertz PD., Spencer SK., Battigelli DA. (2003). Incidence of enteric viruses in groundwater from household wells in Wisconsin. *Appl. Environ. Microbiol.* **69** (2003), pp. 1172–1180

Borchardt MA., Bradbury KR., Alexander EC., Kolberg RJ., Alexander SC., Archer JR. (2010). Norovirus outbreak caused by a new septic system in a dolomite aquifer. *Ground Water* Epub 2010 Feb 22.

Central Statistics Office (CSO) 2012.Irish Census Data 2011 Report. Central Statistics Office, Dublin, Ireland: pp. 1-93.http://www.cso.ie/census/

Charrois JWA. (2010). Private drinking water supplies: challenges for public health. *Canadian Medical Association Journal*. 182(10)

Corkal DR., Schutzman WC., Hilliard CR. (2004). Rural Water Safety from the Source to the On-Farm Tap. *Journal of Toxicology and Environmental Health, Part A.* **67**:1619-1642.

Department of Environment and Local Government. (DoELG) (1999). Groundwater Protection Schemes. Department of Environment and Local Government, Environmental Protection Agency, Geological Survey of Ireland.

Environmental Protection Agency (EPA) (2010). Water Quality in Ireland 2007-2009; Office of Environmental Assessment, EPA. Wexford, Ireland.

Fong TT., Mansfield LS., Wilson DL., Schwab DJ., Molloy SL., Rose JB.(2007). Massive microbial groundwater contamination associated with a waterborne outbreak in Lake Erie, South Bass Island, Ohio. *Environ. Health Perspect*. **115**(6): 856-864

Gaut S. (2005). Factors Influencing Microbiological Quality of Groundwater from Potable Water Supply Wells in Norwegian Crystalline Bedrock Aquifers. PhD thesis, Norwegian University of Science and Technology, Faculty of Engineering Science and Technology.

Geological Survey of Ireland (GSI) (2009). Aquifer classifications in the Republic of Ireland; GSI Draft Document

Giannoulis N., Maipa V., Konstantinou I., Albanis T., Dimoliatis I. (2005). Microbiological risk assessment of Agios Georgios source supplies in north western Greece based on faecal coliform determination and sanitary inspection survey. *Chemosphere.* **58**: 1269-1276.

Godfrey S., Timo F., Smith M. (2006). Microbiological risk assessment and management of shallow groundwater sources in Lichinga, Mozambique. *Wat. Environ. J.* **20**:194–202.

Gonzales TR. (2008) The effects that well depth and wellhead protection have on bacterial contamination of private water wells in the Estes Park Valley, Colorado. *Jour. Env. Health* **71**(5):17-23

Gunnarsdottir MJ., Gardarsson SM., Andradottir HO. (2013). Microbial contamination in groundwater supply in a cold climate and coarse soil: case study of norovirus outbreak at Lake Myvatn, Iceland. *Journal of Hydrology Research* – In Press

Helsel DR., Hirsch RM. (2002). Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 524 p. Available online at *http://water.usgs.gov/pubs/twri/twri4a3/*

Hossain F., Sivakumar B. (2006). Spatial Pattern of Arsenic Contamination in Shallow Tubewells of Bangladesh: Regional Geology and Non-linear Dynamics *Stochastic Environmental Research and Risk Assessment*. **20**(1-2): 66-76

Hosmer DW., Lemeshow S. (2000). Applied Logistic Regression, New York : Wiley

Howard G., Pedley S., Barrett M., Nalubega M., Johal K. (2003). Risk factors contributing to microbiological contamination of shallow groundwater in Kampala, Uganda. *Water Research* **37**, Issue 14, 3421-3429

Hrudey SE., Hrudey EJ. (2007). Published case studies of waterborne disease outbreaks – Evidence of a recurrent threat. *Water Environ. Res.* **79**(3): 233-245

Hynds PD., Misstear BD., Gill LW. (2012) Development of a microbial contamination susceptibility model for private domestic groundwater sources. *Water Resources Research*. **48**(12)

Institute of Geologists of Ireland (IGI) (2007). Guidelines for Drilling Wells for Private Water Supplies.

Lindberg M., Wikström B., Lindberg P. (2010). Subgroups of haemodialysis patients in relation to fluid intake restrictions: a cluster analytical approach. *Journal of Clinical Nursing.* **19**, Issue 21-22, pages 2997–3005, November 2010

Luby SP., Gupta SK., Sheikh MA., Johnston RB., Ram PK., Islam MS. (2008). Tubewell water quality and predictors of contamination in three flood-prone areas in Bangladesh. *Appl. Microbiol*. **105**:1002–1008

Misstear BDR., Daly D. (2000) Groundwater protection in a Celtic region: the Irish example in, editor(s) Robins NS., Misstear BD. *Groundwater in the Celtic regions: Studies in hard rock and karst hydrogeology*, London, Geological Society Special Publication 182, 2000, pp53 – 65.

Misstear BDR, Fitzsimons V (2007) Estimating groundwater recharge in fractured bedrock aquifers in Ireland. Chapter 16 in Groundwater in Fractured Rocks, Special Publication 9 of the International Association of Hydrogeologists, J Krasny & J Sharp (eds), Taylor and Francis, 2007, 243-257

Misstear BDR., Brown L., Daly D. (2009). A methodology for making initial estimates of groundwater recharge from groundwater vulnerability mapping*; Hydrogeology*, 17: 275-285

Nolan B. (2001). Relating nitrogen sources and aquifer susceptibility to nitrate in shallow ground waters of the United States. *Ground Water.* **39**(2): 290–299.

OCM (2007) Establishing Natural Background Levels for Groundwater in Ireland, O'Callaghan Moran & Associates, on behalf of the South Eastern River Basin District Project Team.

Olaniyan IO., Agunwamba JC., Ademiluyi JO. (2010). Assessment of Aquifer Characteristics in Relation to Rural Water Supply in Part of Northern Nigeria**.** *Researcher ;* Vol.2 No.3 , 2010 pp22-27.

Oxfam-DelAgua Ltd., (2004) Oxfam-Delagua Portable Water Testing Kit Users Manual (Version 4.1).

Payne SM., Woessner WW. (2010). An Aquifer Classification System and Geographical Information System-Based Analysis Tool for Watershed Managers in the Western U.S. *JAWRA Journal of the American Water Resources Association*; Volume 46, Issue 5, pages 1003–1023, October 2010

Rozemeijer JC., Broers HP. (2007). The groundwater contribution to surface water contamination in a region with inten-sive agricultural land use (Noord-Brabant, The Netherlands). *Environmental Pollution*. **148**: 695-706

Schijven JF., Mulschlegel JH., Hassanizadeh SM., Teunis PF., de Roda Husman AM. (2006). Determination of protection zones for Dutch groundwater wells against virus contamination – uncertainty and sensitivity analysis. Journal of Water and Health. **04**:297-312

Swartz M., Misstear, BD., Daly D., Farrell, ER. (2003). Assessing subsoil permeability for groundwater vulnerability. *Quarterly Journal of Engineering Geology and Hydrogeology***;** 2003; v. 36; issue.2; p. 173-184

Tabbot P., Robson M. (2006). The New Jersey residential well-testing program – a case study: Randolph township. *Journal of Environmental Health*, **69**(2), 15-19.

Teso RR., Poe MP., Younglove T., McCool PM. (1996). Use of Logistic Regression and GIS Modeling to Predict Groundwater Vulnerability to Pesticides. *Jour. Env. Qual.* 25(3):425-432

U.S. Bureau of the Census. (2008). Current Housing Reports, Series H150/07, American Housing Survey for the United States: 2007, U.S. Government Printing Office, Washington, D.C.: 20401, Printed in 2008. Available athttp://www.census.gov/prod/2008pubs/h150-07.pdf

U.S. Geological Survey (USGS) (2005) National Field Manual for the Collection of Water-Quality Data; Techniques of Water-Resources Investigations Book 9; Handbooks for Water-Resources Investigations.

WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation. (2010). Progress on Sanitation and Drinking Water: 2010 Update. Geneva

Zhang B., Song XF., Zhang YH., Han DM., Tang CY., Yu YL., Ma Y. (2012). Hydrochemical characteristics and water quality assessment of surface water and groundwater in Songnen plain, Northeast China. *Water Res*. **46**:2737-2748.

Figure 1.

Map of Ireland showing location of major towns/cities and study areas

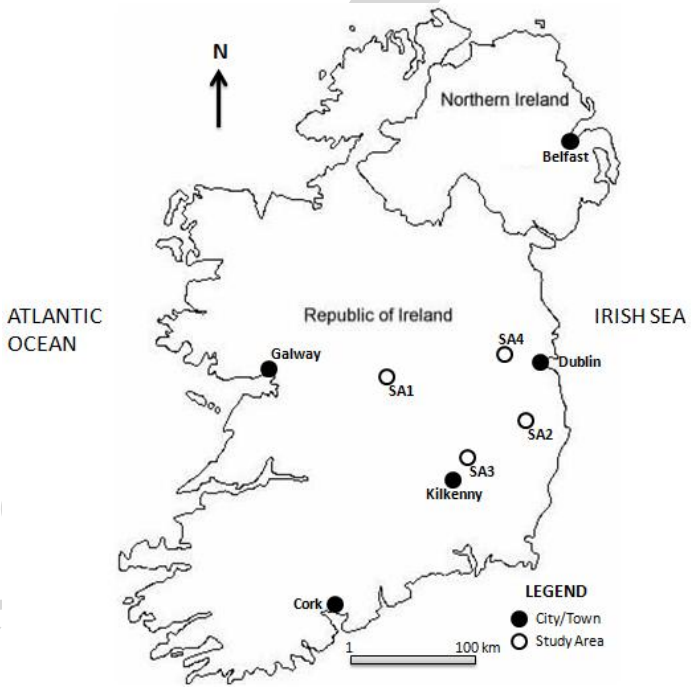

Table 1

Aquifer Classification in the Republic of Ireland

	Regionally Important (R)	Locally Important (L)	Poor (P) Aquifers
Sub-categories	• Karstified where conduit flow is dominant (Rk) • Fractured bedrock aquifers (Rf) • Extensive sand/gravel aquifers (Rg)	• Sand/gravel (Lg) • Generally moderately productive (Lm) • Moderately productively in local zones (Ll) • Karstified (Lk)	• Generally unproductive except for local zones (Pl) • Generally unproductive (Pu)
Criteria for main aquifer categories	- Areal extent >25 km^2 - Well yields >400 $m^3 d^{-1}$ - Specific capacities >40 $m^3 d^{-1} m^{-1}$ - Occurrence of large springs	- Areal extent <25 km^2 - Well yields 100-400 $m^3 d^{-1}$	- Well yields typically <100 $m^3 d^{-1}$ (<40 $m^3 d^{-1}$ in Pu sub-category)

(Adapted from Misstear & Daly, 2000)

Table 2

Study area characteristics including study area size and dominant hydrogeological parameters

Study Area	Area (km^2)	Vulnerability	Dominant Bedrock	Dominant Subsoil	Aquifer Importance*
SA1	15.5	*High*	Sandstone Limestone	Limestone S&Gs Limestone Tills	Ll/Lm
SA2	20.3	*High/Extreme*	Granite	Granite Tills	Pl/Ll
SA3	30.8	*High/Extreme*	Sandstone Shale Limestone	Sandstone, Shale & Limestone Tills	Pu/Lm/Pl
SA4	15.8	*Low*	Limestone Sandstone	Limestone Tills	Ll/Pl

*See Table 1

Table 3

Cluster Centroid Analysis Results and Cluster Characteristics

Cluster	Temp		pH		EC		Boreholes	Hand-dug	*E. coli* (%)
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.			
1	8.11	3.07	6.29	0.62	310.2	156.6	81	0	15.4
2	11.68	2.61	7.11	0.24	775.5	288.3	88	0	34.1
3	10.22	4.12	6.66	0.68	546.6	302.6	0	42	47.6
Combined	10.05	3.5	6.71	0.62	554.8	324.8	169	42	29.4

Table 4

Results of statistical analysis between risk factors and *E. coli* presence (Cluster-1)

Variable	Test Statistic	Sig	df
Point Agricultural Source <100m	8.111[1]	0.006	1
Wellhead Finish	8.996[1]	0.011	2
Liner Clearance	6.883[1]	0.015	1
Septic Tank Setback Distance	2.17[2]	0.033	85
120-hour Precipitation	-2.042[2]	0.047	85

[1] Chi-square tests of independence

[2] Independent samples t-tests

Table 5

Results of statistical analysis between risk factors and *E. coli* presence (Cluster-2)

Variable	Test Statistic	Sig	df
Bedrock Type	20.435[1]	0.009	8
Subsoil Type	12.511[1]	0.006	3
Septic Tank Gradient	12.267[1]	0.002	2
Point Agricultural Source Distance	2.281[2]	0.037	44

[1] Chi-square tests of independence

[2] Independent samples t-tests

Table 6

Results of statistical analysis between risk factors and *E. coli* presence (Cluster-3)

Variable	Test Statistic	Sig	df
Bedrock Type	12.095[1]	0.034	5
Aquifer Importance	8.001[1]	0.046	3
Road Gradient	7.153[1]	0.028	2
Pump Type	8.257[1]	0.016	2
Wellhead Finish	6.011[1]	0.05	2
Houses supplied	2.285[2]	0.032	40
Septic Tank Setback Distance	2.406[2]	0.025	40

[1] Chi-square tests of independence

[2] Independent samples t-test

Table 7

Multivariate model of risk factor hierarchies within Cluster-1

	B	B Sig.	Sig.	H/L Sig.	R^2	Var (%)
Hierarchy 1:						
Septic Tank Setback	-0.075	0.002	0.025	0.39	0.078	17
Hierarchy 2:						
Liner Clearance	-3.095	<0.001	0.002	0.115	0.184	23
Hierarchy 3:						
Point Agricultural Sources <100m	-4.276	0.001	<0.001	0.714	0.356	37
Hierarchy 4:						
Road Gradient (UG/OG)	-2.288	0.009				
Constant	9.194	<0.001	<0.001	0.104	0.463	23

Table 8

Multivariate model of risk factor hierarchies within Cluster-2

	B	B Sig.	Sig.	H/L Sig.	R^2	Var (%)
Hierarchy 1:						
Septic Tank Up Grad	-5.227	0.01				
Septic Tank On Grad	-7.674	0.002	0.009	1	0.193	32
Hierarchy 2:						
Subsoil Thickness	5.755	0.02	0.002	0.64	0.299	18
Hierarchy 3:						
Grazing Land On Grad	-3.168	0.01	<0.001	0.874	0.392	15.5
Hierarchy 4:						
Liner Clearance (mm)	-0.021	0.029				
Above Ground Wellhead Finish	2.89	0.09				
Wellhead Rad. Ground Cond. (10m)	2.273	0.09				
Constant	3.397	0.133	<0.001	0.709	0.597	34

Table 9

Multivariate model of risk factor hierarchies within Cluster-3

	B	B Sig.	Model Sig.	H/L Sig.	R^2	Var (%)
Hierarchy 1:						
Septic Tank Setback	-0.032	0.015				
Septic Tank Up Grad	-5.82	0.021				
Septic Tank Down Grad	-3.63	0.044	0.008	0.862	0.327	42
Hierarchy 2:						
Liner Clearance	-4.478	0.016	0.001	0.545	0.497	21
Hierarchy 3:						
Point Agricultural Sources <100m	4.438	0.019	<0.001	0.811	0.603	14
Hierarchy 4:						
Number of houses supplied	-3.981	0.022				
Constant	13.469	0.008	<0.001	0.543	0.773	20.5

Table 10
Multivariate model prediction accuracy classification

	Observed	Predicted	Percent Correct
Cluster 1			
***E. coli* Absent**	58	54	93.1
***E. coli* Present**	29	19	65.5
			83.9
Cluster 2			
***E. coli* Absent**	68	65	95.6
***E. coli* Present**	12	9	75
			90.9
Cluster 3			
***E. coli* Absent**	22	20	90.9
***E. coli* Present**	20	17	85
			88.1

Groundwater source contamination mechanisms: Physicochemical profile clustering, risk factor analysis and multivariate modeling

Highlights

- Three clusters associated with well susceptibility are identified
- Clusters are associated with well type and geological setting
- Statistical analyses indicate clusters exhibit differing contamination mechanisms
- Multivariate models are used to elucidate well contamination mechanisms