Comparability of macroinvertebrate biomonitoring indices of river health derived from semi-quantitative and quantitative methodologies.

Nicholas C. Everall[1], Matthew F. Johnson[2], Paul Wood[3], Andrew Farmer[1], Robert L. Wilby[3], Nick Measham[4]

1. Aquascience Consultancy Limited, Chesterfield, Derbyshire, S42 7JS
2. School of Geography, University of Nottingham, NG7 2RD, UK
3. Centre for Hydrological and Ecosystem Science, Department of Geography, Loughborough University, Leicestershire, LE11 3TU, UK
4. Salmon and Trout Conservation, UK, Burgate Manor, Fordingbridge, Hampshire

Submitted to: *Ecological Indicators*

Accepted Version

Abstract:

Aquatic macroinvertebrates have been the basis for one of the primary indicators and a cornerstone of lotic biomonitoring for over 40 years. Despite the widespread use of lotic invertebrates in statutory biomonitoring networks, scientific research and citizen science projects, the sampling methodologies employed frequently vary between studies. Routine statutory biomonitoring has historically relied on semi-quantitative sampling methods (timed kick sampling), while much academic research has favoured fully quantitative methods (e.g. Surber sampling). There is an untested assumption that data derived using quantitative and semi-quantitative samples are not comparable for biomonitoring purposes. As a result, data derived from the same site, but using different sampling techniques, have typically not been analysed together or directly compared. Here, we test this assumption by comparing a range of biomonitoring metrics derived from data collected using timed semi-quantitative kick samples and quantitative Surber samples from the same sites simultaneously. In total, 39 pairs of samples from 7 rivers in the UK were compared for two seasons (spring and autumn). We found a strong positive correlation (r_s = +0.84) between estimates of taxa richness based on ten Surber sub-samples and a single kick sample. The majority of biomonitoring metrics were comparable between techniques, although only fully quantitative sampling allows the density of the community (individual m^{-2}) to be determined. However, this advantage needs to be balanced alongside the greater total sampling time and effort associated with the fully quantitative methodology used here. Kick samples did not provide a good estimate of relative abundance of a number of species / taxa and, therefore, the quantitative method has the potential to provide important additional information which may support the interpretation of the biological metrics.

Keywords:

Macroinvertebrate; Species Richness; Biological Monitoring; Biotic Index; River

1. Introduction:

Rivers and the ecological communities they support comprise some of the most biodiverse habitats on the globe but are also some of the most degraded as a result of anthropogenic activity (Dudgeon et al. 2006; Carpenter et al. 2011). River habitats and their ecosystems are threatened by ongoing human development (Vörösmarty *et al*. 2010), including the modification of channel morphology, dredging, changes to catchment land-use, pollution from diffuse and point sources, invasion by alien species, and alterations of the flow regime from abstraction, damming and flood risk management (Carpenter *et al*. 2011). The historic degradation of rivers has prompted the development of a range of biological monitoring tools to survey and quantify anthropogenic stressors over the past 40 years (e.g., Hering et al., 2004) and underpin calls to restore and improve the ecological health of lotic ecosystems (e.g., Geist, 2011).

In order to quantify trends in the health of riverine environments, the response of an organism or community is often characterised as a metric based on their known tolerances to 'stressors'. Biological monitoring, or biomonitoring, can be used to assess the effect of a known change to the state of a system by comparing the ecological community before and after the change or to routinely check compliance to nationally / internationally recognised standards, such as the legal requirement for all waterbodies in the European Union to achieve 'Good Ecological Status' under the Water Framework Directive. The taxonomic resolution of such indices varies from family-level metrics that give broad indications of water quality (e.g., Walley and Hawkes 1997) to species/genus-level metrics that can provide information about specific stressors (Hubler et al., 2016); although some can be used at different taxonomic resolutions (Monk et al., 2012). Other metrics use higher resolutions; for example, the phenology of species or groups of species can be used to assess the impacts of climate change (Everall et al. 2015; Thackeray et al. 2016).

Aquatic macroinvertebrates are a fundamental component of freshwater ecosystems. Hence, maintaining macroinvertebrate communities, biodiversity and individual species populations contributes to the overall ecological integrity of the system (Spänhoff and Arle, 2007). Particular invertebrates (species, genus or families) have tolerance limits to specific environmental conditions, such as levels of salinity, pH, organic pollution, suspended sediment concentration, fine sediment deposition and flow velocity (e.g. Hellawell, 1986). Macroinvertebrate biomonitoring tools and assessment systems are widely used to assess water quality globally (e.g. North America – Barbour et al. 1999; Africa – Cummins *et al*. 2004; Asia – Morse *et al*. 2007; South America – Dickens & Graham, 2002), although there have been recent calls for methods of assessing ecological response to environmental changes and pressures to be more strongly rooted in ecological and biological theory (e.g. Friberg *et al*. 2011; Johnson and Rice, 2014). In Europe, macroinvertebrate biomonitoring forms an important part of compliance monitoring within the European Union Water Framework Directive (WFD). This Directive requires Member States to ensure that all freshwater bodies are of 'Good Ecological Status (GES) or Good Ecological Potential (GEP) for Heavily Modified Waterbodies (HMWB) and Artificial Waterbodies (AWB) by 2027 (EU Directive 2000/60/EC).

Biomonitoring techniques can be quantitative, semi-quantitative or qualitative, depending on the technique used. The most common method for sampling invertebrates in rivers is the semi-quantitative kick sample method, where invertebrates are sampled over a specified time period (typically three-minutes) supplemented by hand searches of larger substrate clasts; although the total area or proportion of the community sampled is typically unknown (Murray-Bligh, 1999; ISO 10870, 2012). Most macroinvertebrate biomonitoring indices have been developed to allow macroinvertebrate community composition to be analysed on a semi-quantitative basis where sampling effort (time) is standardised (Clements and Newman, 2002). Fully quantitative sampling is necessary for other forms of analysis that require information regarding the total abundance, density or diversity of organisms/communities within a specified area. This can be achieved with a Surber sampler (or other similar devices such as a cylinder sampler, or Hess sampler), where invertebrates are collected within a specified sampling area.

Whilst there is widespread agreement that the macroinvertebrate community provides a valuable tool to characterise the ecological health of rivers, there is less consensus about the most appropriate sampling methodologies to employ. Surprisingly, the degree to which biological metrics derived from semi-quantitative and quantitative samples differ has not been widely assessed in a systematic way. The largely untested assumption that biomonitoring scores are not comparable between these methods prevents both historic (e.g. Percival and Whitehead, 1929; Percival and Whitehead, 1930; Prigg, 2002) and contemporary fully quantitative data from being combined and used to characterise river health. Hence, the aim of this paper is to compare a semi-quantitative kick sampling methodology with a quantitative Surber sampling methodology at given sites by cross-matching: 1) derived biomonitoring scores/indices; 2) inferred water- and habitat-quality; and 3) the abundance and diversity of the taxa collected by each method.

2. Methodology:

2.1. Sampling techniques

Kick sampling is a semi-quantitative method of surveying the invertebrate community, which is widely used internationally because it is cost effective and results are relatively consistent between operators (e.g. Carter and Resh, 2001; Metzeling *et al*., 2003). In this study, a 1 mm^2 mesh net with an opening 0.25 m wide and 0.22 m deep was held downstream of the operator who kicked the river bed and swept the net through, for example, submerged macrophytes. This action disturbs sediment and dislodges benthic invertebrates which are then carried by the river flow into the net. The duration of kick sampling here followed the Environment Agency of England (EA) best-practice standard, which requires three-minutes of kick sampling and one-minute hand search of larger substrates for macro-invertebrates (HMSO, 1985, Murray-Bligh, 1999; Environment Agency, 2009). The operator moved systematically across and upstream through the river reach being sampled, ensuring that all main habitat types were sampled (e.g. emergent and submerged macrophyte stands, woody debris, tree roots, different flow depth/velocities and bed substrate compositions). The amount of time spent in each designated habitat unit was proportionate to the surface area that each occupied.

To obtain a quantitative comparison, replicate Surber samples were collected. A Surber sampler is a rectangular quadrat, 0.33 x 0.30 m (area 0.1 m^2) that is placed on the river bed. The quadrat has a 1 mm^2 mesh net attached, with a 0.29 x 0.34 m opening. The operator disturbs by hand all surface material within the quadrat area. Total sampling times can vary but in the current study continued until all of the 0.1 m^2 quadrat area was fully sampled (Surber, 1937; Macan, 1958). Sediment was disturbed to a maximum depth of 0.1 m. Disturbance dislodges invertebrates that then drift into the downstream net and, with the aid of side curtains, captures dislodged animals that might otherwise avoid capture in the net. Traditional Surber net sampling tended to be micro-habitat specific but for some river types Surber net sampling can form part of a methodology that proportionally samples different microhabitats (Prigg, 2002; Everall, 2010). In this study, 10 Surber samples, distributed such that all habitat types within the site were represented, were undertaken at each survey site. As with kick sampling, the habitats sampled reflected the proportion of the area covered by each habitat type at the site. For ease of analysis, the 10 individual samples were aggregated into 5 sub-samples for identification. The data from these 5 sub-sample units were, in turn, aggregated prior to the calculation of the biomonitoring indices/scores used for comparison between methods.

All samples were collected following the EA best practice guides (Environment Agency, 2009) by an experienced operator (Everall). Kick and Surber sampling was undertaken on the same day, at the same site, one immediately after the other. The second sample was taken a few metres upstream of the first but spatially alternating between kick and Surber net sample reaches at survey sites to reduce any sampling sequence bias. Sample site reaches were selected for their similarity of instream habitat composition over the sampled reach and were divided into kick and Surber areas such that each had comparable proportions of the major habitat types.

2.2. Sampling times and locations

Sampling was undertaken on seven English rivers at a total of 20 sites (Figure 1). These locations were chosen to provide a range of habitat and climate types (Table 1). Geology and elevation were obtained from Ordnance Survey maps. Average discharge and average annual maximum discharge were derived from daily average and daily maximum discharge time-series from the nearest gauging station on each river available from the Centre for Ecology and Hydrology (CEH). The 1961-1990 average annual precipitation for the area upstream of gauging stations is also included in Table 1.

Kick and Surber samples were undertaken in spring (March-May) and autumn (September-October) at all sites on all rivers except for the River Wye where a kick and Surber sample pair was only taken in spring (Table 1). In total, 39 paired kick and Surber samples were collected. All samples were identified by the same laboratory technician to species level where possible. Where not possible, invertebrates were identified to the highest possible taxonomic level.

2.2. Biological scoring methods

A set of ecological parameters and biological monitoring scores were calculated for each site (Table 2). These represent commonly applied metrics in the UK that are used to identify water quality and more specific environmental stressors. The abundance and taxa richness of the whole community was quantified, as well as the diversity of Ephemeroptera, Plecoptera and Trichoptera (EPT) and *Gammarus*, which are important sentinels of environmental stressors in the UK. The abundance or proportion of EPT taxa is widely used and considered to be a good indicator of river health where salmonid fisheries are economically important (Stanford and Spacie, 1994; Clements and Newman, 2002; Park *et al*. 2003). In addition, the Community Conservation Index (CCI; Chadd and Extence; 2004) provides an indication of exceptionally rich or regionally unusual invertebrate populations by scoring invertebrates based on their rarity. The CCI can contribute to the overall description of the condition of an aquatic ecosystem, alongside indices designed to detect, for example, flow variation or changes in water quality.

The Biological Monitoring Working Party (BMWP) score, ranks individual macroinvertebrate families from 1 to 10 based on their sensitivity to water quality. The sum of the scores of all collected families is the BMWP score. Given that the BMWP score is affected by the number of families sampled but not by abundances within those families, the interpretation can be biased as a sample with many low scoring taxa might score the same as a sample with a few high scoring taxa. Therefore, the Average Score Per Taxon (ASPT) was introduced, where the BWMP is divided by the total number of scoring families, to provide an average measure (Armitage et al. 1983). The Whalley Hawkes Paisley Trigg (WHPT) biometric score (Paisley et al. 2013) was developed as an attempt to integrate the abundance weighting limitation of the BMWP scoring system. These are indicative of family-level aggregate and averaged biomonitoring scores and are part of the WFD assessment criteria in the UK, with similar systems implemented across Europe.

Stressor-specific indicators were also deployed. The saprobic index is used to assess organic pollution by assigning a value (the saprobic value, *s*) to each invertebrate species or family which indicates their tolerance to organic pollution. Each invertebrate is also given an indicator value (*G*), that represents the tolerance range of an invertebrate and acts as a weighting value, increasing the impact of very sensitive organisms on the overall saprobic score (S). All saprobic values were obtained from Schmidt-Kloiber and Hering (2015b). The Saprobic indicator was used here because it is internationally recognised and used as a good indicator of organic enrichment and pollution and it was the forerunner for many contemporary systems.

Other stressor-specific indices used here include the Proportion of Sediment-sensitive Invertebrates (PSI), Lotic-invertebrate Index for Flow Evaluation (LIFE) and Total Reactive Phosphorous Index (TRPI). Both the PSI and LIFE score are regularly applied in the UK, particularly to sites that are in danger of not achieving WFD requirements. The PSI is used to assess the presence of fine sediment by calculating the percentage of sediment sensitive taxa present in a sample (Extence et al. 2010). Similarly, the LIFE score uses the proportion of flow sensitive invertebrates in a sample to describe the prevailing flow conditions at that site (Extence et al. 1999). Finally, the TRPI (Everall, 2010) uses the proportion of phosphorous tolerant and intolerant macroinvertebrates in a sample according to various river types and seasons (Paisley et al., 2003; Paisley *et. al.*, 2011). These scores are good examples of classification systems were the percentage or proportion of sensitive organisms are compared to the total community.

2.3. Interpretation

To enable interpretation of the indices a ranking system was used, where 1 indicates poor conditions (highly stressed/impacted conditions) and 5 indicates very good conditions (un-stressed and non-impacted conditions) (Table 3). The scoring system used herein is based on established knowledge where available (see references in Table 3). Biomonitoring scores were grouped into each of these classes and the differences in grouping between kick and Surber sample results were compared. The WHPT score is interpreted using the River Invertebrate Classification Tool (RICT), a software program that compares observed WHPT scores to expected scores (see Paisley et al. 2007; UKTAG, 2014) and therefore simple categorisation is not appropriate for this metric. Given that all the metrics are continuous and judgement is necessary for data that fall near the boundary of a class, the difference between kick and Surber samples as a percentage of the category size was also determined. This indicates the likelihood that a methodological difference would lead to the results falling into a different category.

Where a biomonitoring score has an inconsistent range within categories the average class size was calculated. For example, in the case of the BMWP, the middle condition (rank 3) has a range of 19 whereas good (rank 4) has a range of 24. Therefore, it is possible for a difference between kick and Surber sampling to be greater than 100% of a class size but with both samples actually being in the same category. In addition, where both kick and Surber samples are in the highest category, it is possible to achieve scores that differ by more than 100% of a class boundary but within the same class because there is not a higher category.

2.4. Statistics

The statistical significance of differences between sets of biological scores calculated with kick and Surber sampled data were tested. Shapiro-Wilk tests indicated data was normally distributed with the exception of the total abundance, abundance of *Gammarus*, species richness, CCI and EPT diversity. Paired-sample Student t-tests were performed in SPSS v.22 to assess normally distributed data. In the case of non-normally distributed data, a Wilcoxon Signed Rank test was performed instead. In addition, Pearson correlation and linear regression analysis was used to compare normally distributed kick and Surber sampled data. Where data was not normally distributed, Spearman correlation applied (r_s). Initially, this was performed for each biological monitoring score, incorporating data collected at all sites and seasons (n = 39). The data are spatially clustered and in some instances comprise multiple samples from the same site at different times of year. However, the regression analysis was not describing relationships between sites or times of year, but between sampling strategies. Therefore, the clustering of data does not affect the robustness of the test. If the null hypothesis is met and both sampling methods provide identical information, the R^2 should equal 1 and the data should fall on the 1:1 line (i.e. y = x). Subsequently, linear regression analysis was also performed on spring and autumn data, separately, in a sub-set of cases.

3. Results:

3.1. Invertebrate abundance, diversity and community measures

In total, 128,129 individual invertebrates were sampled across all sites and techniques (78 samples), representing 205 different taxa. At sites where Surber samples collected a high abundance of invertebrates, the equivalent kick sample also tended to collect a high relative abundance. Hence, the relationship between kick and Surber samples was significantly positively correlated ($r_s = +0.64$; $p < 0.001$). However, there was considerable scatter in the association (Figure 2).

In 90% of the samples, the total number of invertebrates collected was higher in the aggregated Surber samples than in the kick samples. Similarly, the total number of EPT collected was greater in Surber samples than equivalent kick samples in 85% of cases. The abundance of *Gammarus* sp. in samples was more similar between sampling methods, with only 62% of sites having greater abundance in Surber samples. Where Surber samples collected a greater abundance than the paired kick sample, they contained, on average, twice as many invertebrates as the equivalent kick sample. In contrast, the kick samples that were more abundant than Surber samples yielded, on average, only 1.2 times more individuals than the paired Surber samples (Table 4). The total invertebrate abundance and total EPT abundance for kick and Surber samples were significantly different ($p < 0.01$, in both cases). The total number of *Gammarus* sp. sampled did not differ statistically between sampling techniques ($p = 0.062$).

The total diversity of invertebrates collected in Surber samples was positively correlated with the taxa richness of equivalent kick samples ($r_s = +0.84$, $p < 0.001$). Correlations for taxa richness were stronger than for measures of abundance, but there was still considerable scatter (Figure 3a, b). In general samples collected following the Surber sample methodology were more taxa rich than equivalent kick samples, with 70% of samples having more taxa in the Surber than the kick. The difference in species richness and EPT richness between kick and equivalent Surber samples was statistically significant in both cases ($p < 0.001$, in both cases).

The CCI calculated from Surber and kick net samples are positively correlated (r_s = +0.81; p < 0.001) and are statistically similar (p = 0.499) (Figure 3c), indicating similarity in the collection of rarer taxa between methods.

3.2. Biomonitoring scores

Paired-sample Student t-tests indicate that the differences between the BMWP, ASPT and WHPT calculated from kick and Surber sampled data were not statistically different for any metric (p = 0.06; p = 0.955, p = 0.08, respectively). BMWP, ASPT and WHPT displayed strong, statistically significant correlations between Surber and kick sampled results (BWMP r = +0.85, p < 0.001; ASPT r = +0.88, p < 0.001; WHPT r = +0.93, p < 0.001). There was scatter in each relationship, but slightly more variance was explained for WHPT (R^2 = 0.87) than for the ASPT (R^2 = 0.78) and BMWP (R^2 = 0.74) (Figure 4).

The difference between each of the four stress-sensitive metrics when calculated on Surber and kick sampled data were statistically indistinguishable (Saprobic p = 0.656; TRPI p = 0.147; PSI p = 0.143; LIFE p = 0.166) (Figure 5). All four metrics showed a strong relationship between Surber and kick sampled data, and were all significantly positively corrected (p < 0.001 in all cases). The strongest association between kick and Surber sampled data was for the PSI and LIFE scores, both of which are based on the proportion of sensitive invertebrates to all sampled invertebrates.

The TRPI score displayed the lowest R^2 of the stress specific metrics, although the R^2 = 0.78 still suggests a strong relationship between kick and Surber sampled results. The TRPI was affected by two outliers where the Surber sample scored 100% whereas the equivalent kick sample scored substantially less. When these two outliers were removed, R^2 increases to 0.90.

Comparing kick and Surber methods taken in the spring with those collected during the autumn indicated that spring samples were generally more consistent between sampling methods (Table 5). There was more variation between the two sampling methods in autumn for all biological metrics, with the exception of the BMWP, ASPT and Saprobic index, which were slightly more consistent in the autumn.

3.3. Score interpretation

Differences between biomonitoring scores calculated on Surber and kick sampled data are sufficient to alter the resulting classification of 35 (15%) of the biometric scores (Table 6). In 17 cases, the kick samples returned a higher class category than the Surber sample method, whereas the reverse was true in 18 cases. On average, the BMWP calculated using the Surber sample methodology was 63% of a class boundary greater than the kick sampled equivalent. The ASPT differed by an average of 22% of a class boundary and the saprobic index by 15% of a class boundary.

The LIFE score differed by 19% and the PSI by 19% of a class boundary and the equivalent value for the TRPI was 23% (Table 7). In general, kick samples returned higher ranking of the PSI and Saprobic Index. As the scoring systems were continuous, rankings could be altered by small increments in score if they fall close to the class boundary. To assess the likelihood that a difference in sampling method would lead to different class interpretation, the difference between kick and Surber sample methodology scores was presented as a percentage of the number within each class (Table 7).

3.4. Preferential sampling of particular species

Across all aggregated sites, some species of invertebrate were consistently more likely to be caught using the Surber sample than by the equivalent kick sample method and, to a lesser extent, the opposite was observed for a small number of taxa. Some invertebrates, such as *Gammarus pulex* and *Baetis* sp., were recorded at much greater abundances in the Surber sample than the kick sample method (Figure 6). For example, nearly twice as many *Agapetus* sp. caddisfly and three-times as many Simuliid blackfly larvae were found in total across all Surber samples. In contrast, kick samples caught more *Limnephilus lunatus* (cased caddisfly larvae) and the amphipod shrimp *Crangonyx pseudogracilis* than equivalent Surber samples (Figure 6). Whilst more abundant, these invertebrate taxa were not found at more sites and, consequently, any sampling bias did not alter biological metrics between methods. However, some invertebrates were found at more sites, with potential implications for biomonitoring scores (Figure 7). Notable examples were the bivalve *Pisidium* sp. and the caseless caddisfly *Lype reducta* which were both recorded in more kick samples than equivalent Surber samples (6 and 5 more sites, respectively; Figure 7). In contrast, the leeches *Piscicola geometra* and *Helobdella stagnalis* were recorded in 8 more Surber samples than kick samples. There were 27 taxa only recorded in Surber samples in contrast to 7 taxa only found in kick samples (Supplementary A). Those only found in kick samples were only ever observed at one site whereas some of the invertebrates only recorded in Surber samples were sometimes found at multiple sites.

4. Discussion:

4.1. Sensitivity of biomonitoring scores to sampling method

Differences in the invertebrate community collected using the Surber and kick sample methods affect the biomonitoring scores that are derived to varying degrees and levels of significance. The BMWP was most affected, because this is calculated by aggregating the score associated with identified families. Hence, any increased diversity of Surber samples leads to higher BMWP scores. The effect of different sample sizes was reduced to some extent by the ASPT score, which was more similar between sampling methods. The WHPT was most consistent, with this method both averaging the score by the total abundance, as well as standardising invertebrate scores by individual family abundance within the sample.

Stress-specific scores were similar for data collected via Surber and kick sampling. Kick and Surber sample LIFE and PSI scores were both highly correlated and statistically similar. However, for the sites sampled here, the PSI was consistently higher for the kick sample, probably because the kick sample was not as effective at collecting sediment-dwelling invertebrates which tend to reduce the score. One explanation for this is that hand disturbance of surface grains and the aim to disturb sediment to 10 cm depth in Surber samples is likely to dislodge more subsurface material. Furthermore, the Surber net has a wider mouth for sample collection and hand sampling causes less hydrodynamic disturbance than kick sampling (which may drive some animals around the net entrance). The Surber net also has retention sides or curtains at the mouth to aid sample capture which the kick-sweep net does not.

The saprobic and TRPI were also consistent between kick and Surber sampling, although the latter was affected by an outliers. These are important findings for the Saprobic index since loss and gain of species numbers have indicated a strong mechanistic link with elevated and declining organic enrichment (BOD levels) across UK rivers with differing Surber and kick sampling techniques employed in recent years (Clews and Ormerod, 2009; Durance and Ormerod, 2009; Everall, 2010).

4.2. Sensitivity of water- and habitat-quality to sampling method

Variance between biomonitoring scores calculated with kick and Surber sampled data can lead to different interpretations if a ranking classification is used. In the current investigation, all scores differed on average by less than a single class, although the BMWP does differ on average by 63% of a class boundary. However, this was largely associated with very high scoring Surber samples where the equivalent kick sample was already in the top class. Hence, the BMWP was actually the metric where boundary classifications were most consistent between the two methods examine. The least consistent was the PSI, despite being very highly correlated, statistically similar and with an average difference of only 19% of a class boundary. This is likely because many of the sites fell close to class boundaries and even a slight but consistent bias in kick sampled data was sufficient to under-represent sediment dwelling invertebrates.

Although not explicitly tested here, it is likely that the uncertainty due to the choice of Surber or kick sampling method is comparable to the uncertainty when comparing between different operators, at different times of year, in different areas. For example, there are natural seasonal variations in all biometrics because of temporal changes in macroinvertebrate community structure, life cycle stages and the concomitant response of the seasonally resident invertebrate communities to the ephemeral toxicity of contaminants (Hynes, 1970; Hellawell, 1989; Clements and Newman, 2002). Overall, metrics were more similar between kick and Surber samples in spring but this was dependent on the biomonitoring score used. Data presented here suggest that the difference in metrics at the same site between spring and autumn can be twice as great as the difference when comparing metrics collected using kick and Surber sampling techniques. This underlines the importance of sampling across known natural variations in invertebrate community structures and seasonal impacts of pollution to fully characterise water quality.

Previous research shows that inherent uncertainties in sampling and identifying macroinvertebrate samples can substantially exceed those described here, associated with sample collection. For example, Haase *et al*. (2010) audited river macroinvertebrate biomonitoring as part of an EU WFD requirement. A subset of samples processed by government agencies were re-processed by auditors who found that 29% of specimens and 21% of all taxon were overlooked when sorting and that individuals successfully selected in the sorting processes were correlated to body-size. Over 30% of taxa were identified differently between individuals and auditors, which was not biased towards harder to identify individuals. As a result of these differences, 34% of samples were categorised into a different quality classes. Similarly, Carter and Resh (2001) found in the USA that different methods of data collection, sub-sampling and sorting were commonly used yet these were known to yield different results. Here, leeches and flatworms were recorded preferentially when using the Surber sampler method which could be because of more limited detritus present in Surber samples, making these animals easier to distinguish than in the paired kick samples.

4.3. Sensitivity of invertebrate community to sampling method

The Surber sample method collected significantly more invertebrates (abundance) and a significantly greater diversity of invertebrate species than the kick sample method, both in spring and autumn. This is particularly true of the EPT taxa. For example, the Surber sample method collected twice as many *Ephemera danica* mayfly larvae when aggregated across all samples than equivalent kick samples (Figure 6). Similarly, invertebrates that attach themselves to the sediment were more prevalent in samples using the Surber sampler method (e.g. Simuliidae blackfly larvae) (Figure 6). This was expected given the increased sampling effort when compared to the three-minute kick sample method. The kick samples were limited to 3-minutes but Surber samples continued until all the surface area had been disturbed, resulting in a longer overall sampling time than kick samples.

Invertebrates that were found preferentially by one method over the other will potentially alter biomonitoring scores. An example is the cased caddisfly larvae, *Glossosoma* sp., which were recorded at seven sites using Surber sampling in comparison to only two kick sample sites. Other organisms more likely to be recorded using the Surber sampler than the kick sample method included the leeches *Helobdella stagnalis* and *Piscicola geometra* which were found in 15 and 17 Surber samples, but only 5 and 10 kick samples, respectively. Similarly, the flatworm *Polycelis felina* was found in six more Surber samples than equivalent kick samples. It may be that these sediment-dwelling animals are caught more efficiently in Surber samples where sampling is attempted to a depth of 10 cm, ensuring that sub-surface material is thoroughly disturbed.

The only two organisms identified that were consistently observed in more kick samples than Surber samples, was the caseless caddisfly *Lype reducta*, which was found in seven of the 39 kick samples in comparison to only two of the equivalent number of Surber samples, and individuals in the bivalve genus *Pisidium*, which were found in 6 more kick samples than Surber samples. The reason for this is not clear, but in the case of *Lype reducta* it could possibly be because they are xylophagous and have a close association with coarse wood on the river bed.

These results are consistent with the study of Gillies *et al*. (2009) who found kick samples collected only 63% of taxa that were collected in quantitative Surber samples in New South Wales, Australia. Gillies *et al*. (2009) also found that kick samples were biased towards sampling large, abundant and widely distributed taxa, with those missed generally being smaller in size and rarer in the wider environment. In the current study, individual samples using the kick sample method were not obviously biased towards larger species, because even large invertebrates such as *Ephemera danica* (body length > 20 mm) were under-represented in kick samples. Similarly, Storey et al. (1991) found that Surber and kick samples in south-western Australia were broadly similar, but with key differences represented by Sorensen's similarity coefficients of 66% in June and 61% in September.

4.4. Added value of a quantitative sample

There is a great deal of data held in records that have been used to generate biomonitoring scores, which could provide additional, valuable information. However, where qualitative or semi-quantitative measures have been used, the comparability of data is not readily assessed given the lack of information about the proportion of the river bed or invertebrate population that has been sampled. Although kick samples here did generally under-represent some taxa, they did provide a sufficiently good estimate of the invertebrate diversity to provide statistically similar biomonitoring scores to the quantitative Surber sample. However, the kick sample did not provide a good estimate of the relative abundance of many species. Given that this information is not required for many biomonitoring scores, this does not affect the value of biological metrics calculated. However, without a good estimate of total abundance, it is difficult to make ecological assertions about the community. In addition, not quantifying the abundance of taxa may lead to loss of important information, such as changing abundance / occurrence through time which may be indicative of a chronic issue but which would not be identified by most biomonitoring scores unless species are also concurrently impacted from the community. The Surber sampling method used here provides a quantitative measure of population (e.g. the abundance / m^2), so it provides added value over semi-quantitative methods, allowing a more thorough investigation of the data, which may lend support or add detail to the information gained from the use of biological metrics.

5. Conclusions:

This study set out to establish the extent to which community, biomonitoring scores, and inferred environmental conditions, are sensitive to the choice of invertebrate sampling method. Our analysis was based on an English data set covering 20 sites, 205 taxa and 128,129 identified organisms. We found that the overall taxa richness of aquatic invertebrates that were collected in quantitative Surber samples were greater than semi-quantitative kick sample equivalents, although the two were positively correlated. Surber samples enable additional ecological information and analysis to be undertaken and, at least at the sites studied here, gave a more complete overview of the abundance and diversity of macroinvertebrates. However, biomonitoring scores did not differ significantly in most cases and, therefore, a semi-quantitative kick sample methodology provided a suitable estimate of the river health of the chosen sites. In particular, specific pressure based biomonitoring scores which use an abundance weighting (ratio of sensitive to total invertebrate abundance), such as the LIFE, PSI and TRPI scores, yielded very similar results, regardless of the sampling method.

The comparability of biometric indices from Surber and kick-sweep net sampling raises the possibility of using historical Surber net sample data to assess longer-term trends in biological stress signatures. Based upon the findings here, a wider use of replicated Surber net sampling is proposed, particularly where it is necessary to detect rare taxa that may be endangered or for 'one-off' quantitative and statistically testable benchmarking of ecological condition in river reaches, additional to routine regulatory monitoring programmes.

Acknowledgements

We are grateful to Salmon and Trout Conservation UK for contributing towards the funding of this research. Discharge data was obtained from the UK National River Flow Archive. We would also like to thank two anonymous reviewers for their helpful comments.

6. Reference:

Armitage, P.D. Moss, D. Wright, J.F. & Furse, M.T. (1983) The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. *Water Research* **17**: 333–347.

Barbour M.T., Gerritsen J., Snyder B.D. & Stribling J.B. (1999) *Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish.* 2nd edition. EPA 841-B-99±002. U.S. Environmental Protection Agency; Office of Water; Washington D.C, USA.

Bonada, N., Prat, N., Resh, V.H. and Statzner, B. (2006) Developments in aquatic insect biomonitoring: A comparative analysis of recent approaches. *Annual Review of Entomology* **51**: 495–523.

Carpenter, S.R., Stanley, E.H. and Vander Zanden, M.J. (2011) State of the World's freshwater ecosystems: Physical, chemical and biological changes. *Annual Review of Environment and Resources* **36**: 75–99.

Carter, J.L. and Resh, V.H. (2001) After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *Journal of the North American Benthological Society* **20**: 658–682.

Chadd, R. and Extence, C. (2004). The conservation of freshwater macroinvertebrate populations: a community-based classification scheme. *Aquatic Conservation: Marine and Freshwater Ecosystems* **14**, 597–624.

Clements, W.H. and Newman, M.C. (2002). *Community Ecotoxicology.* Hierarchical Ecotoxicology Series. John Wiley and Sons Ltd, Hoboken, USA.

Clews, E. and Ormerod, S.J. (2009). Improving bio-diagnostic monitoring using simple combinations of standard biotic indices. *River Research and Applications* **25**, 348–261.

Cummins, K.W., Merritt, R.W. and Andrare P.C.N. (2004) The use of functional groups to characterise ecosystem attributes in selected streams and rivers in south Brazil. *Studies on Neotropical Fauna and Environment* **40**: 69–89.

Dickens C.W.S. and Graham, P.M. (2002) The South African Scoring System (SASS) Version 5 rapid bioassessment method for rivers. *African Journal of Aquatic Science* **27**: 1–10.

Dufrene, M. and Legendre, P. (1997) Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs* **67**: 345–366.

Dudgeon, D., Arthington, A.H., Gessner, M.O., Kawabata, Z., Knowler, D.J., Lévêque, C., Naiman, R.J. Prieur-Richard, A-H, Soto, D., Stiassny, M.L.J. and Sullivan, C.A. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews* **81**: 163–182.

Durance, I. and Ormerod, S.J. (2009). Trends in water quality and discharge confound long-term warming effects on river macroinvertebrates. *Freshwater Biology* **54**: 388–405.

Environment Agency (2009) *Freshwater Macro-invertebrate sampling in Rivers*. Operational Instruction 018_08. Environment Agency, Bristol.

Everall, N.C., Johnson, M.F., Wilby, R.L. and Bennett, C.J. (2015) Detecting phenology change in the mayfly *Ephemera danica*: Responses to spatial and temporal water temperature variation. *Ecological Entomology* **40**: 95–105.

Everall, N.C. (2010). *The aquatic ecological status of the rivers of the Upper Dove Catchment in 2009. Natural England Commissioned Report NECR046*. Natural England: Sheffield.

Extence C.A., Balbi, D.M. and Chadd R.P. (1999). River Flow Indexing using British benthic macroinvertebrates: A framework for setting hydro ecological objectives. *Regulated Rivers Research and Management* **15**, 543–574.

Extence, C.A, Chadd, R.P, England, J., Dunbar, M.J., Taylor, E.D. and Everall, N.C. (2010). The assessment of fine sediment accumulation in rivers using macroinvertebrate community response. BHS Third International Symposium, Managing Consequences of a Changing Global Environment, Newcastle, 2010.

Extence, C.A., Chadd, R.P., England, J., Dunbar, M.J., Wood, P.J. and Taylor, E.D. (2013). The Assessment of Fine Sediment Accumulation in Rivers Using Macro-invertebrate Community Response. *River Research and Applications* **29**: 17–55.

Eyre, M., Pilkington, J.G., McBlane, R.P. and Rushton, S.P. (2005). Macroinvertebrate species and assemblages in the headwater streams of the River Tyne, northern England, in relation to land cover and other variables. *Hydrobiologia* **544**, 229–240.

Friberg, N., Bonada, N., Bradley, D.C., Dunbar, M.J., Edwards, F.K., Grey, J., Hayes, R.B., Hildrew, A.G., Lamouroux, N., Trimmer, M. and Woodward, G. (2011) Biomonitoring of human impacts in freshwater ecosystems: the good, the bad and the ugly. *Advances in Ecological Research* **44**: 1 – 68.

Geist, J. (2011) Integrative freshwater ecology and biodiversity conservation. *Ecological Indicators* **11**: 1507–1516.

Gillies, C.L., Hose, G.C. and Turak, E. (2009) What do qualitative rapid assessment collections of macroinvertebrates represent? A comparison with extensive quantitative sampling. *Environmental Monitoring and Assessment* **149**: 99–112.

Haase, P., Pauls, S. Schindehütte, K. and Sundermann, A. (2010) First audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program: human error greatly lowers precision of assessment results. *Journal of the North American Benthological Society* **29**: 1279–1291.

Hawkins C. P., Yong C., Roper B. (2010). Method of predicting reference condition biota affects the performance and interpretation of ecological indices. *Freshwater Biology* **55**: 1066–1085.

Hellawell, J.M. (1986). *Biological indicators of freshwater pollution and environmental management*. Pollution Monitoring Series. Elsevier Applied Science, London, UK.

Hering, D., Moog, O., Sandin, L., Verdonschot. P.F. (2004) Overview and application of the AQEM assessment system. *Hydrobiologia*, **516**: 1–20

H.M.S.O. (1985). *Methods of Biological Sampling Handnet Sampling of Aquatic Benthic Macroinvertebrates*. Methods for the Examination of Waters and Associated Materials. London, UK.

ISO 10870 (2012) Water quality -- Guidelines for the selection of sampling methods and devices for benthic macroinvertebrates in fresh waters.

Hubler, S., Huff, D.D., Edwards, P. and Pan, Y. (2016). The Biological Sediment Tolerance Index: Assessing fine sediments conditions in Oregon streams using macroinvertebrates. *Ecological Indicators* **67**: 132–145.

Hynes, H.B.N. (1970) *The Ecology of Running Waters*. Liverpool University Press, Liverpool, UK.

Johnson, M.F. and Rice, S.P. (2014) Animal perception in gravel-bed rivers: Scales of sensing and environmental controls on sensory information. *Canadian Journal of Fisheries and Aquatic Sciences* **71**: 945–957.

Macan, T.T. (1958). Methods of sampling the bottom fauna in stony streams. *Mitteilung Internationale Vereinigung fuer Theoretische unde Amgewandte Limnologie* **8**: 1–21.

Metzeling, L., Chessman, B., Hardwick, R. and Wong, V. (2003) Rapid assessment of rivers using macroinvertebrates: the role of experience, and comparisons with quantitative methods. *Hydrobiologia* **510**: 39–52.

Monk, W.A., Wood, P.J., Hannah, D.M., Extence, C.A., Chadd, R.P., and Dunbar, M.J. (2012). How does macroinvertebrate taxonomic resolution influence hydroecological relationships in riverine ecosystems. *Ecohydrology* **5**: 36–45.

Morse, J.C., Bae, Y.J., Munkhjargal, G., Sangpradub, N., Tanida, K., Vshivkova, T.S., Wand, B., Yang, L and Yule, C.M. (2007) Freshwater biomonitoring with macroinvertebrates in East Asia. *Frontiers in Ecology and the Environment* **5**: 33–42.

Murray-Bligh, J. A. D. (1999) *Procedure for collecting and analysing macro-invertebrate samples.* Quality Management Systems for Environmental Biology: Biological Techniques, BT001 Version 2.0. Bristol: Environment Agency, UK.

Paisley, M.F., Walley, W.J., Nikhade, J. and Dils, R. (2003). Identification of the key biological indicators of nutrient enrichment in rivers for use in predictive/diagnostic models. Proceeding of the 7th International Specialised IWA Conference on Diffuse Pollution and Basin Management, Dublin, Ireland.

Paisley, M.F., Trigg, D.J. and W.J. Walley (2007) Revision and Testing of BMWP scores. Final report SNIFFER Project WFD72a. Edinburgh, SNIFFER.

Paisley, M.F., Walley, W.J. and Trigg, D.J. (2011) Identification of macro-invertebrate taxa as indicators of nutrient enrichment in rivers. *Ecological Informatics* **6**, 399–406.

Paisley, M.F., Walley, W.J. and Trigg, D.J. (2013) Revision of the Biological Monitoring Working Party (BMWP) score system: Derivation of present-only and abundance related scores from field data. *River Research and Applications* **30**: 887–904.

Park, Y-S., Cereghino, R., Compin, A. and Lek, S. (2003). Application of artificial neaural networks for patterning and predicting aquatic species richness in running waters. *Ecological Modelling* **160**: 265–280.

Pantle R. and Buck H. (1955) Die biologische Uberwachung der Gewas und die Darstellung der Ergebnisse. *Gas und Wasserfach* **96**: 603.

Percival, E. and Whitehead, H. (1929). A quantitative study of some types of stream bed. *Journal of Ecology* **17**: 282–314.

Percival, E. and Whitehead, H. (1930). Biological survey of the River Wharfe II. Report on the invertebrate fauna. *Ibid* **18**: 286–302.

Phillips, D.J.H. and Rainbow, P.S. (1993) *Biomonitoring of trace aquatic contaminants*. Elsevier Applied Science, New York, USA.

Prigg, R.F. (2002) *The Eden Fly Life Project: Examining Indications of Abundance of Riverine Aquatic Insects over Several Decades*. Environment Agency, Technical Memorandum NB 729(6/03).

Rapport, D.J., Reiger, H.A. and Hutchinson, T.C. (1985) Ecosystem behaviour under stress. *American Nature* **125**, 617–640.

Schmidt-Kloiber, A. and Hering, D. (2015a) www.freshwaterecology.info – an online tool that unifies, standardises and codifies more than 20,000 European freshwater organisms and their ecological preferences. *Ecological Indicators* **53**: 271–282.

Schmidt-Kloiber, A. and Hering, D. (2015b; eds) www.freshwaterecology.info version 6.0 (Accessed on 16/02/2017).

Spänhoff, B. and Arle, J. (2007). Setting attainable goals of stream habitat restoration from a macroinvertebrate view. *Restoration Ecology* **15**, 317–20.

Stanford, L.L. and Spacie, A. (1994). *Biological Monitoring of Aquatic Systems*. CRC Press, Florida, USA.

Storey, A.W., Edwards, D.H.D. and Gazey, E.P. (1991) Surber and kick sampling: a comparison for the assessment of macroinvertebrate community structure in streams of south-western Australia. *Hydrobiologia* **211**: 111–121.

Surber, E.W. (1937) Rainbow trout and bottom fauna production in one mile of stream. *Transactions of the American Fisheries Society* **66**: 193–202.

Thackeray, S.J., Henrys, P.A., Hemming, D., Bell, J.R., Botham, M.S., Burthe, S., Helaouet, P., Johns, D.G., Jones, I.D., Leech, D.I., Mackay, E.B., Massimino, D., Atkinson, S., Bacon, P.J., Beretin, T.M., Carvalho, L., Clutton-Brock, T.H., Duck, C., Edwards, M., Elliot, J.M., Hall, S.J.G., Harringtno, R., Pearce-Higgins, J.W., Høye, T.T., Kruuk, L.E.B., Pemberton, J.M., Sparks, T.H., Thompson, P.N., Hite, I., Winfield, I.J. and Wanless, S. (2016) Phenological sensitivity to climate across taxa and trophic levels. *Nature* **535**: 241–245.

UK TAG (2014) Guide to Invertebrates in Rivers. Invertebrates (General Degradation): Whalley, Hakes, Paisley & Trigg (WHPT) metric in River Invertebrate Classification Tool (RICT). WFD – UK Technical Advisory Group Method Statement.

Vörösmarty, C.J., McIntrye, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, E., Sullivan, C.A., Reidy Liermann, C. and Davies, P.M. (2010) Global threats to human water security and river biodiversity. *Nature* **467**: 555–561.

Walley W. J., Grbovic, J., and Dzeroski, S. (2001) A reappraisal of saprobic values and indicator weights based on Slovenian river quality data. *Water Research* **35**: 4285–4292.

Walley, W.J. and H.A. Hawkes (1997) A computer-based reappraisal of the Biological Monitoring Working Party score system incorporating abundance rating, site type and indicator value. *Water Research* **31**: 201–210.

WFD (2015) Water Environment (Water Framework Directive) (England and Wales) (Amendment) Regulations 2015.

Wright J.F., Sutcliffe D.W. and Furse, M.T. (2000). *Assessing the Biological Quality of Fresh Waters. RIVPACS and Other Techniques*. Freshwater Biological Association, Ambleside, UK.

Zelinka, M. and Marvan, P. (1961) Zur Präzisierung der biologischen Klassifikation der Reinheit fliessender Gewässer. *Archiv für Hydrobiologie* **57**: 389–407.

Figure 1: A map of England and Wales with the 7 sampled rivers with circles.

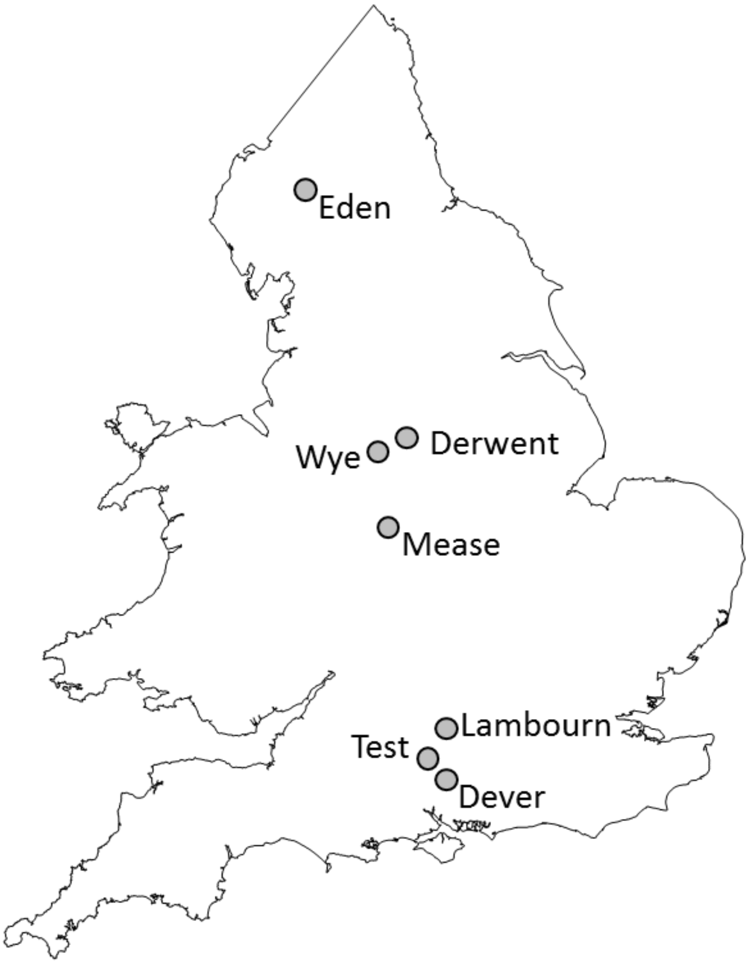


Figure 2: Relationship between the a) total invertebrate abundance, b) total EPT and c) total *Gammarus* collected in Surber samples versus kick samples, taken on the same day and at the same site. Pearson (r) and Spearman (r_s) correlation coefficients are included for normal and non-parametric data, respectively.

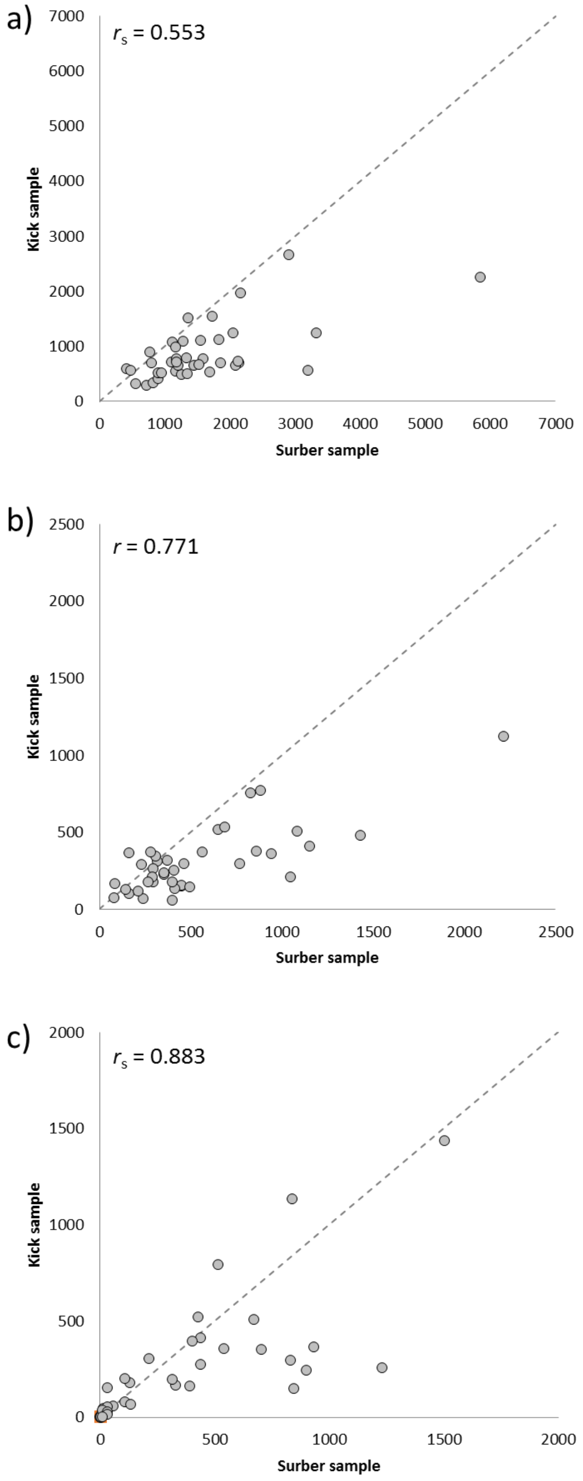


Figure 3: Relationship between a) taxa richness, b) EPT richness (e.g. mayfly, stonefly and caddisfly; EPT) and c) the CCI collected in Surber and kick samples, taken on the same day and at the same site. Spearman (r_s) correlation coefficients are included.

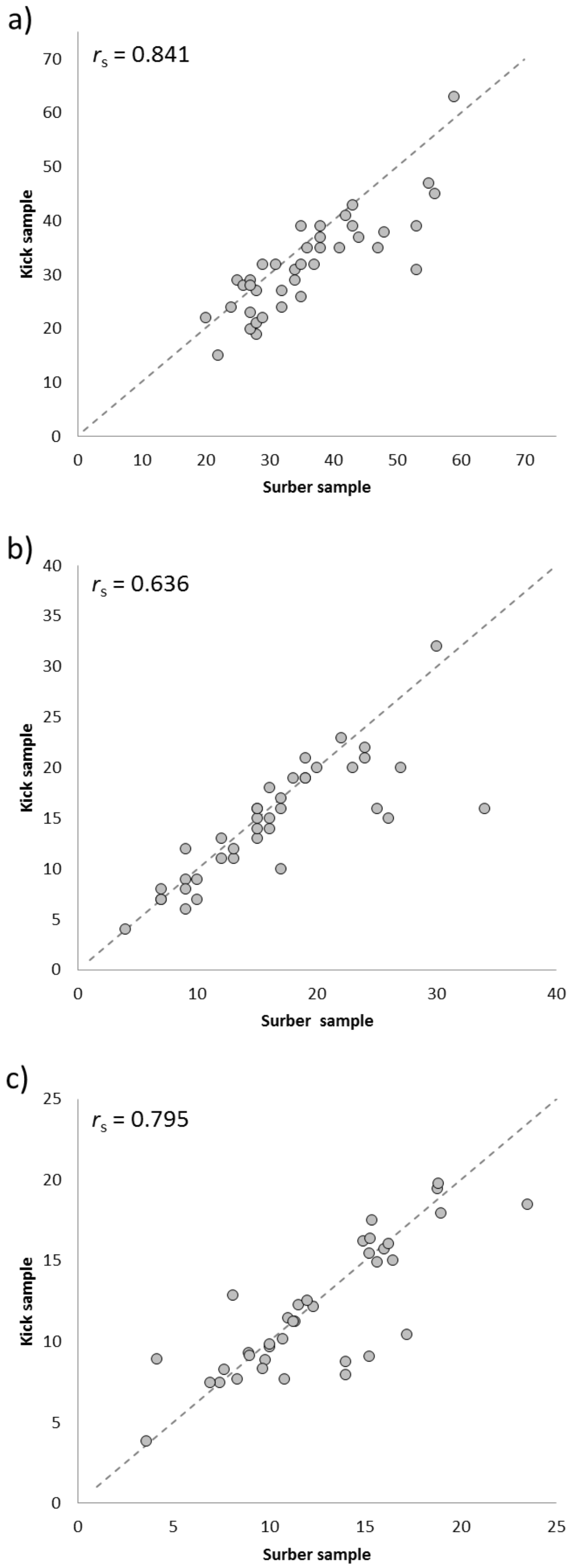
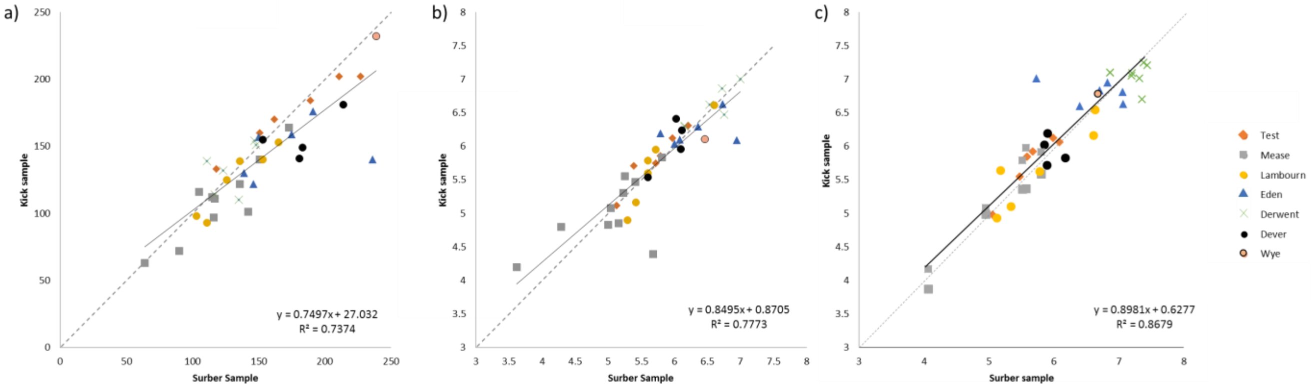
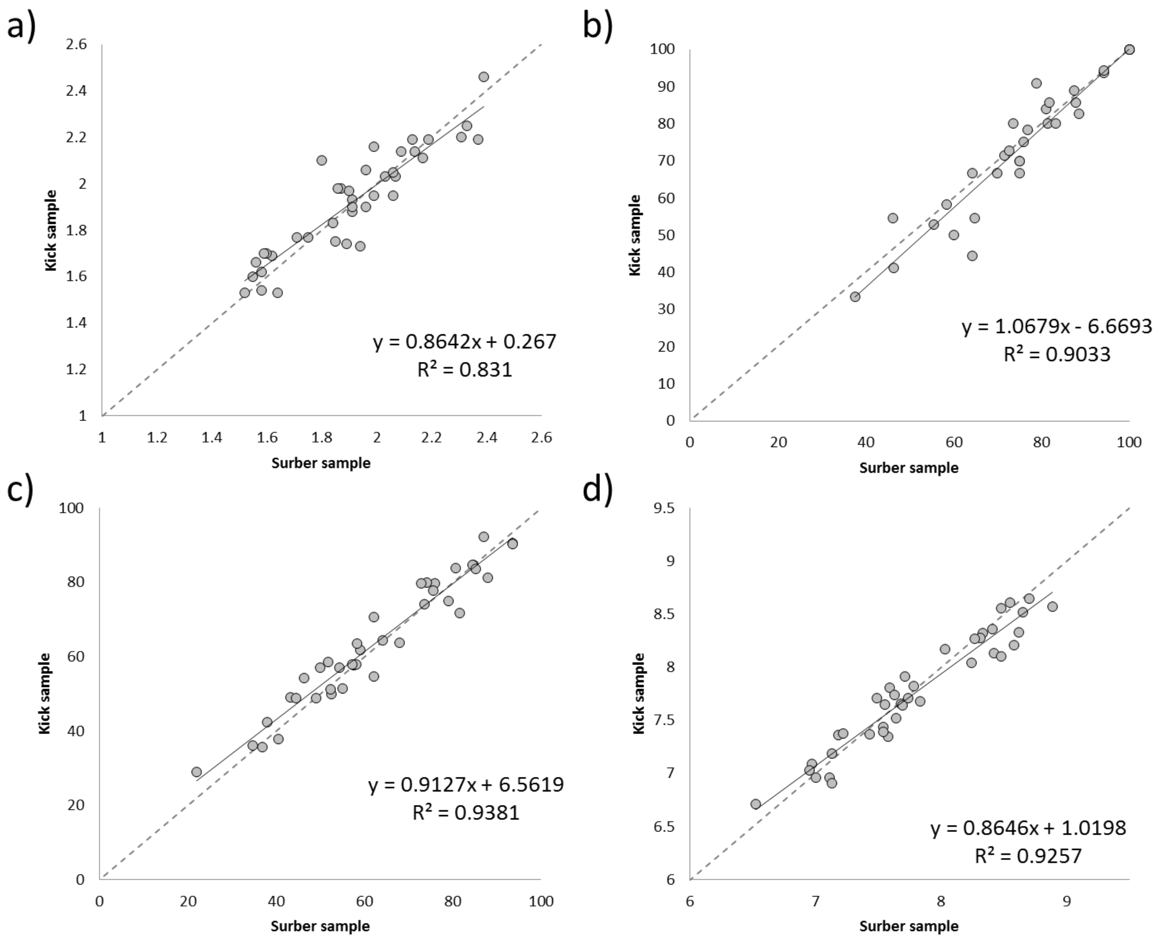


Figure 4: Relationships between the (a) BMWP, (b) ASPT and (c) WHPT calculated using data from Surber versus kick samples, taken on the same day and at the same site. Points are colour-coded to designate the river where the sample was taken.



645 **Figure 5:** Relationships between a) the Saprobic index, b) the TRPI, c) PSI and d) LIFE score calculated

646 on Surber and kick samples, taken on the same day and at the same site.



647

Figure 6: The difference in abundance between kick and Surber samples for invertebrate taxa, aggregated across all sites. Note that the right-hand grey dashed line marks a transition in the horizontal axis from categorical values to absolute values. Taxa of note due to large differences between kick and Surber samples are labelled. Note that in some cases taxa were grouped to genus level (e.g. Baetis sp.) because differences in the proportion of individuals successfully identified to species level (as opposed to genus level) could otherwise have biased results.

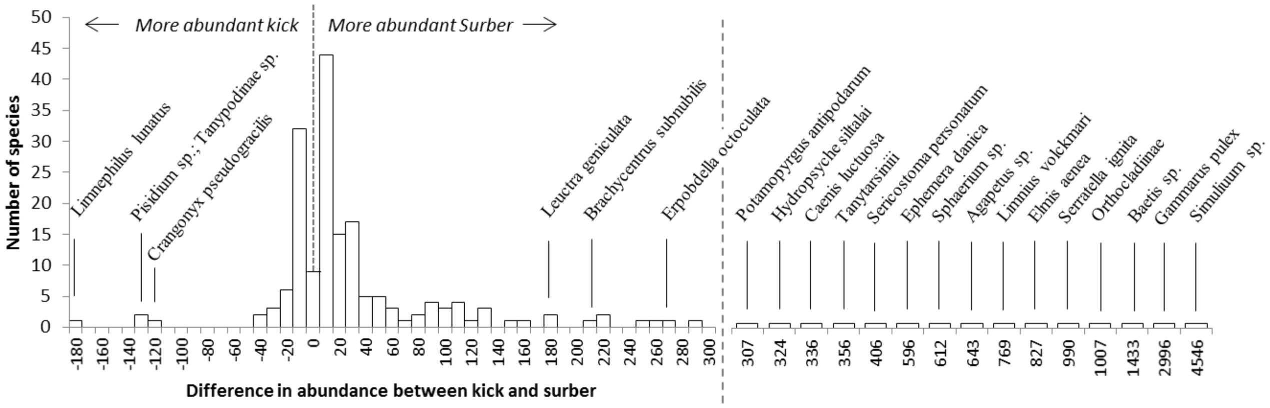


Figure 7: The difference in the number of sites where taxa were caught between kick and Surber samples. Taxa that were found at four or more additional sites for one method are labelled. Zero indicates the taxa was found in the same number of kick and Surber samples.

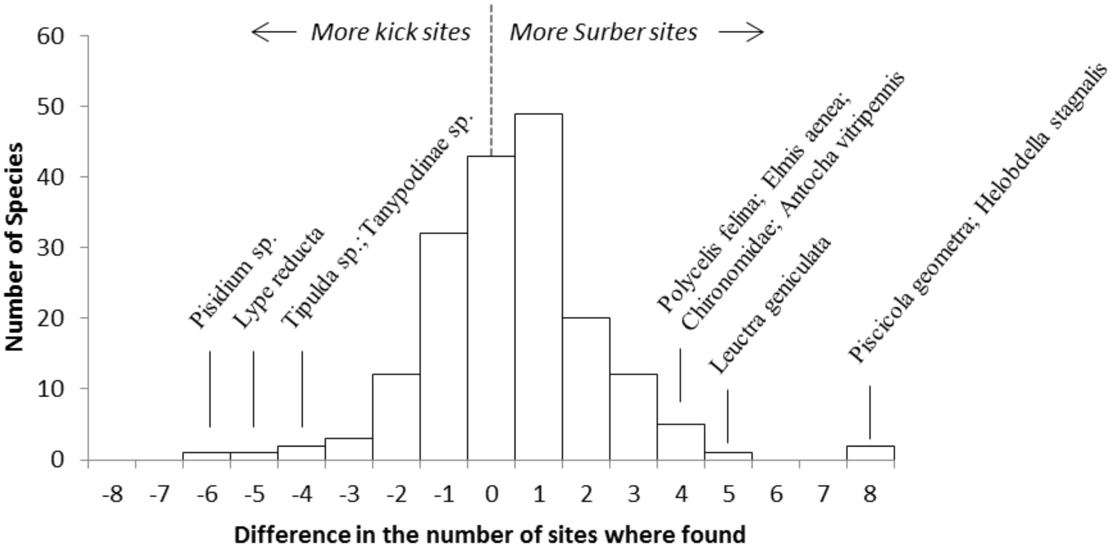


Table 1: *The dates and locations of sampling sites with representative geographic, climatic and hydrologic information for the 7 rivers studied. Land cover proportions were derived from LCM2007 imagery; precipitation information is taken from the UK Met Office 30 year average and discharge information is derived from a 44 year record of gauged flow from the National River Flow Archive.*

	Derwent	Dever	Eden	Lambourn	Mease	Test	Wye
Number of Sites	3	2	3	3	5	3	1
Grid Ref	SK 24671 74452	SU 43300 41999	NY 55831 36050	SU 43371 70208	SK 22166 11370	SU 34838 21355	SK 24367 65787
Date: Spring	19/04/2015	24/04/2015	24/04/2015	14/04/2015	17/05/2013	05/03/2013	22/05/2013
Date: Autumn	14/10/2015	29/09/2015	09/09/2015	01/10/2015	12/09/2013	24/09/2013	
Upstream catchment (km^2)	203	122	616	176	167	453	154
Geology	*Carboniferous sandstone*	*Cretaceous Chalk*	*Permian & Triassic Sandstones*	*Cretaceous Chalk*	*Triassic sandstone/ Mercia mudstone*	*Cretaceous chalk / Paleogene clay*	*Carboniferous Mudstone*
Arable / Grassland (%)	53	57	81	52		46	84
Woodland cover (%)	10	10	5	9		15	4
Urban cover (%)	0.2	0.5	0.4	0.4		1.6	2.3
Site elevation (masl)	139	50	92	96		10.1	139
Annual Precipitation (mm)	1325	780	1146	745		790	1166
Average discharge (m^3 s^{-1})	5.0	1.11	15.0	1.04	N/A	11.2	1.0
Q$_{10}$ (m^3 s^{-1})	11.4	2.0	34.8	1.8	N/A	17.5	6.2

Table 2: *Definitions of ecological parameters and biological monitoring scores tested.*

Parameter	**Definition**
Community Parameters	
Total abundance (A)	The total number of all collected invertebrate taxa
Total diversity (*R*)	The total number / richness of taxa collected
EPT abundance	The total number/ abundance of all collected Ephemeroptera; Plecoptera, Trichoptera taxa
EPT diversity	The total number / richness of all Ephemeroptera; Plecoptera, Trichoptera taxa
Gammarus abundance	The total number of all shrimp (*Gammarus* sp.) collected
Community Conservation Index (CCI)	The national and regional rarity and therefore conservation value of the species community profile
Water Framework Directive Assessment Tools	
Biological Monitoring Working Party Score (BMWP)	The BMWP score calculated with family-level data. No metric for species level
Average Score Per Taxon (ASPT)	The ASPT calculated with family-level data. No metric for species level. It is the BMWP divided by the number of scoring families
Whalley Hawkes Paisley Trigg (WHPT) method	The WHPT is calculated with family-level data. No metric for species level. It uses BMWP scoring system, but scores are dependent on abundance of each scoring family.
Specific Stressor Indicators	
Saprobic Index (S)	The weighted average sensitivity of the invertebrate species community to organic pollution
Proportion of Sediment-sensitive Invertebrates (PSI)	The proportion of sediment-sensitive invertebrates at species level (PSI)
Lotic–invertebrate Index for Flow Evaluation (LIFE)	The proportion of flow-sensitive invertebrates at species level (LIFE)
Total Reactive Phosphorous Index (TRPI)	The proportion of invertebrates sensitive to total reactive phosphorous impact at family level (TRPI)

Table 3: *Class rankings for each biological parameter used, where 1 indicates highly impacted/polluted conditions and 5 indicates un-impacted conditions. An indication of whether the classification is based on the authors' judgement or established knowledge is also given.*

Rank	**1 (v. poor)**	**2**	**3**	**4**	**5 (v. good)**	
A	≤ 99	*100 - 249*	*250 - 999*	*1000 – 4999*	*> 5000*	*Judgement*
R	*< 14*	*15 - 24*	*25 - 34*	*35 – 44*	*> 44*	*Judgement*
EPT	≤ 1	*2 - 9*	*10 - 19*	*20 – 29*	*> 30*	*Judgement*
CCI	*0 - 5*	*5 - 10*	*10 - 15*	*15 – 20*	*> 20*	*Chadd and Extence 2004*
BMWP	*0 - 35*	*36 - 50*	*51- 70*	*71 – 95*	*> 96*	*Hellawell, 1986*
ASPT	*< 5*	*< 5*	*5 - 6*	*6 – 6.5*	*> 6.5*	*Hellawell, 1986; Wright et al. 2000*
S	*3.2 - 4*	*2.7 - 3.19*	*2.3 – 2.69*	*1.81 – 2.29*	*1.0 – 1.8*	Schmidt-Kloiber and Hering 2015b
PSI	*0 - 20*	*21 - 40*	*41 - 60*	*61 - 80*	*81 – 100*	*Extence et al. 2011*
LIFE	*< 6*	*6 – 6.49*	*6.5 – 6.99*	*7 – 7.99*	*> 8*	*Extence et al. 1999*
TRPI	*0 - 20*	*21 - 40*	*41 - 60*	*61 - 80*	*81 – 100*	*Everall 2010*

Table 4*: The number of sites where either Surber or kick samples were more abundant in terms of total abundance,* Gammarus *abundance, and EPT abundance. The average, maximum and minimum difference in abundance, between Surber samples and kick samples is also given.*

	Surber samples more abundant			**Kick samples more abundant**		
	Total (A)	***Gammarus***	**EPT**	**Total (A)**	***Gammarus***	**EPT**
Average	2.08	2.27	2.17	1.22	2.14	1.50
Max	5.71	5.68	6.73	1.41	4.88	2.27
Min	1.03	1.02	1.10	1.11	1.07	1.00
n	36	24	33	5	15	7

Table 5: *The gradient, intercept and amount of explained variance (R^2) by linear regression between biological monitoring scores derived from Surber and kick samples when performed on spring data, autumn data, and spring combined with autumn data. All regressions were significant at* p *< 0.01. Associated graphs can be seen as Supplementary Material B.*

Score	**Season**	**Gradient**	**Intercept**	**R^2**
BMWP	*Spring*	0.944	23.615	0.60
	Autumn	0.929	14.88	0.79
ASPT	*Spring*	0.828	1.062	0.71
	Autumn	0.975	0.089	0.86
WHPT	*Spring*	0.929	0.376	0.93
	Autumn	0.854	0.924	0.80
PSI	*Spring*	1.062	6.712	0.97
	Autumn	0.984	1.459	0.90
LIFE	*Spring*	1.095	0.749	0.95
	Autumn	1.069	0.445	0.90
Saprobic	*Spring*	0.936	0.132	0.80
	Autumn	1.041	0.105	0.89
TRPI	*Spring*	1.095	0.749	0.95
	Autumn	0.996	7.011	0.63

Table 6: *Number of cases where differences in biomonitoring score calculated using kick and Surber samples results in that site being assigned to a different class. A score of 1 indicates the kick sample is one class above the equivalent Surber and -1 indicates the kick sample is one class below the equivalent Surber. The table also shows the difference in biomonitoring score as a percentage of the average class boundary. Values are shaded when the percentage difference is more than 50% of a class boundary. All sites on all rivers are included for samples taken in spring (Sp) and autumn (Au).*

River	Site and Season		Number of classes different						% difference of class boundaries					
			BWMP	ASPT	PSI	S	TRPI	LIFE	BWMP	ASPT	PSI	S	TRPI	LIFE
Derwent	1	Sp	0	0	0	0	0	0	28	0	-14.6	14	1.6	-6.7
	1	Au	0	-1	0	0	0	0	-100	-28	0.8	10	0	-17.3
	2	Sp	0	0	0	0	0	0	112	9	-16.2	-22	0	-42.7
	2	Au	0	0	0	0	0	0	-12	14	1.5	8	0	-38.7
	3	Sp	0	0	0	0	0	0	12	14	26.8	20	-1.9	10.7
	3	Au	0	0	0	0	0	0	36	14	-33.8	20	0	-49.3
Dever	2	Sp	0	0	0	0	0	0	-132	13	-20.5	4	-5	-4.0
	2	Au	0	0	0	0	-1	0	-136	-14	14.0	-2	6.7	5.3
	3	Sp	0	-1	1	0	0	0	8	-6	0.3	-8	-41.7	-16.0
	3	Au	0	0	0	0	0	0	-160	38	42.4	-6	7.0	26.7
Eden	1	Sp	0	0	0	1	0	0	-64	-10	18.6	-30	-10.9	-38.7
	1	Au	0	0	0	0	-1	0	-96	2	-20.1	34	-50.8	-50.7
	2	Sp	0	-1	0	1	1	0	-384	-85	29.5	-42	60.3	-1.3
	2	Au	0	1	0	0	0	0	-36	40	33.9	0	11.9	18.7
	6	Sp	0	0	0	0	-1	0	-60	-7	2.2	-8	-7.4	-6.7
	6	Au	0	0	0	0	0	0	28	4	10.9	4	-25	-4.0
Lambourn	1	Sp	0	0	1	1	0	0	12	19	25.7	-20	19.5	14.7
	1	Au	0	0	0	0	0	0	-20	-26	3.8	-12	-25	13.3
	2	Sp	0	0	0	0	0	0	-4	23	1.5	22	-0.6	-20.0
	2	Au	-1	0	0	0	-1	0	-72	-39	-18.0	23	-99.3	-20.0
	3	Sp	0	0	0	0	0	0	-48	1	16.4	2	0	8.0
	3	Au	0	0	1	0	-1	0	-52	0	-49.5	14	-183.3	-26.7
Mease	1	Sp	0	0	0	0	0	0	-36	6	29.6	-8	-13.1	24.0
	1	Au	0	0	0	0	0	0	-44	2	-12.7	10	0	-8.0
	2	Sp	0	0	0	0	0	0	-164	-129	-0.7	0	31.8	-20.0
	2	Au	0	-1	0	0	0	-1	44	-17	-37.0	20	0	-30.7
	3	Sp	0	0	0	1	0	0	-72	51	35.2	-16	42.0	25.3
	3	Au	0	0	0	1	0	0	-4	58	39.8	-22	0	8.0
	4	Sp	-1	-1	1	0	0	1	-76	-31	22.6	14	-20.9	16.0
	4	Au	0	0	-1	0	0	0	-24	30	-6.3	0	0	-5.3
	5	Sp	0	0	0	1	0	1	-56	4	6.8	-36	-25.8	10.7
	5	Au	0	0	-1	0	0	0	-12	7	22.5	02	0	21.3
Test	1	Sp	0	1	1	0	0	0	-100	15	33.3	22	14.4	29.3
	1	Au	0	0	0	0	0	0	32	7	-5.3	-12	0	-13.3
	2	Sp	0	0	1	0	0	0	36	32	35.7	-22	-29.2	29.3
	2	Au	0	0	0	0	0	0	-20	3	-1.3	-2	-16.7	-2.7
	3	Sp	0	0	1	-1	-1	0	-36	10	14.0	60	-16.7	-6.7
	3	Au	0	0	-1	0	0	-1	60	-1	-13.5	12	-50.0	-29.3
Wye	1	Sp	0	0	0	0	0	0	-28	-35	-7.2	12	-62.5	0

Table 7: *Percentage difference between samples taken in spring and autumn, using both a kick and Surber method. The percentage difference between kick and Surber samples in spring and kick and Surber samples in autumn are also shown.*

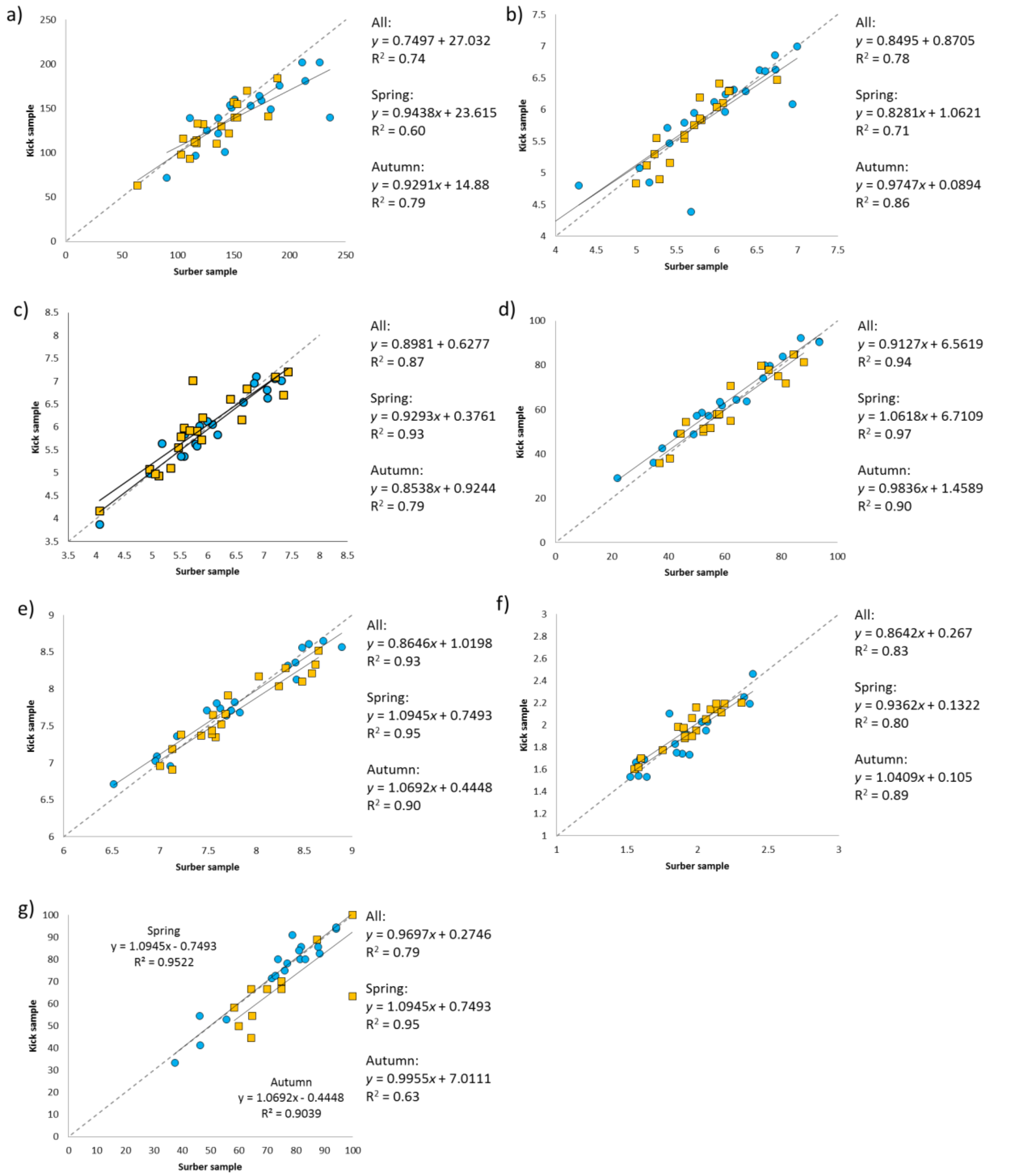
	Total Abundance	EPT	Gammarus	R	EPT	BMWP	ASPT	WHPT	PSI	Sap	TRPI	LIFE
Between spring and autumn												
Kick	71.7	137.6	704.4	25.3	37.7	20.8	9.1	6.4	14.6	8.9	19.9	3.4
Surber	46.3	88.8	709.4	29.2	36.7	26.4	9.0	5.3	11.8	7.8	46.3	3.0
Between kick and Surber												
Spring	99.4	107.3	81.0	19.9	20.5	15.3	4.7	3.4	6.7	5.1	5.0	1.6
Autumn	95.1	99.1	103.0	16.4	15.0	9.1	3.4	4.6	6.1	2.9	12.7	2.0

Supplementary Material A: *Taxa that preferentially occur in either kick or Surber samples. The difference in the number of samples between kick and Surber samples is presented, along with the percentage difference between kick and Surber samples. Only those taxa where the percentage difference is >50% are included.*

Phylum/ Class	Order	Family	Species name	% difference between kick and surber	Difference in number of samples
Insecta	Coleoptera	Dytiscidae	*Agabus didymus*	100% in kick	1
Insecta	Coleoptera	Dytiscidae	*Scirtes* sp.	100% in kick	1
Insecta	Diptera	Muscidae	*Limnophora* sp.	100% in kick	1
Insecta	Diptera	Ptychopteridae		100% in kick	1
Annelida	Arhynchobdellida	Erpobdellidae	*Erpobdella testacea*	100% in kick	1
Platyhelminthes	Tricladida	Planariidae	*Polycelis tenuis*	100% in kick	1
Crustacea	Decapoda	Astacidae	*Austropotamobius pallipes*	100% in kick	1
Insecta	Trichoptera	Psychomyiidae	*Lype reducta*	71% in kick	5
Mollusca	Veneroida	Sphaeriidae	*Pisidium* sp.	50% in kick	6
Insecta	Plecoptera	Perlidae	*Dinocras cephalotes*	100% in Surber	1
Insecta	Ephemeroptera	Baetidae	*Centroptilum luteolum*	100% in Surber	1
Insecta	Ephemeroptera	Heptageniidae	*Ecydonurus dispar*	100% in Surber	1
Insecta	Ephemeroptera	Ephemeridae	*Ephemera vulgata*	100% in Surber	1
Insecta	Trichoptera	Hydroptilidae	*Agraylea multipunctata*	100% in Surber	2
Insecta	Trichoptera	Leptoceridae	*Ceraclea nigronervosa*	100% in Surber	1
Insecta	Trichoptera	Glossosomatidae	*Glossosoma* spp.	100% in Surber	3
Insecta	Trichoptera	Limnephilidae	*Hydatophylax infumatus*	100% in Surber	1
Insecta	Trichoptera	Limnephilidae	*Limnephilus marmoratus*	100% in Surber	1
Insecta	Trichoptera	Leptoceridae	*Mystacides azurea*	100% in Surber	1
Insecta	Trichoptera	Hydroptilidae	*Oxyethira* spp.	100% in Surber	1
Insecta	Trichoptera	Phryganeidae	*Phryganea grandis*	100% in Surber	1
Insecta	Trichoptera	Limnephilidae	*Potamophylax* spp.	100% in Surber	1
Insecta	Trichoptera	Leptoceridae	*Ylodes conspersus*	100% in Surber	1
Insecta	Trichoptera	Hydropsychidae	*Hydropsyche angustipennis*	100% in Surber	1
Insecta	Trichoptera	Hydropsychidae	*Hydropsyche contubernalis*	100% in Surber	2
Insecta	Trichoptera	Hydropsychidae	*Hydropsyche* sp.	100% in Surber	2
Insecta	Trichoptera	Polycentropodidae	*Plectrocnemia conspersa*	100% in Surber	2
Insecta	Trichoptera	Polycentropodidae	*Polycentropus irroratus*	100% in Surber	1
Insecta	Coleoptera	Haliplidae	*Brychius elevatus*	100% in Surber	2
Insecta	Coleoptera	Haliplidae	*Haliplus* spp.	100% in Surber	1
Insecta	Diptera	Empididae	*Chelifera* sp.	100% in Surber	1
Insecta	Diptera	Ptychopteridae	*Ptychoptera* sp.	100% in Surber	2
Mollusca		Bithyniidae	*Bithynia leachi*	100% in Surber	1
Mollusca		Planorbidae	*Planorbis carinatus*	100% in Surber	1
Mollusca	Veneroida	Sphaeriidae	*Pisidium nitidium*	100% in Surber	2
Insecta	Plectoptera	Leuctridae	*Leuctra geniculata*	83% in Surber	5
Insecta	Diptera	Muscidae	*Lispe* spp.	80% in Surber	4
Insecta	Diptera	Chironomidae		67% in Surber	4
Annelida	Rhynchobdellida	Glossiphoniidae	*Helobdella stagnalis*	67% in Surber	8

Annelida	Rhynchobdellida	Piscicolidae	*Piscicola geometra*	53% in Surber	8

Supplementary Material B: *Linear regressions of Surber versus kick samples, collected on the same day and at the same site for a) BMWP, b) ASPT, c) WHPT, d) PSI, e) LIFE, f) Saprobic and g) TRPI scores. Blue circles were taken in spring and orange squares in autumn.*


a)
Kick sample
Surber sample
All:
y = 0.7497 + 27.032
R² = 0.74
Spring:
y = 0.9438x + 23.615
R² = 0.60
Autumn:
y = 0.9291x + 14.88
R² = 0.79
b)
All:
y = 0.8495 + 0.8705
R² = 0.78
Spring:
y = 0.8281x + 1.0621
R² = 0.71
Autumn:
y = 0.9747x + 0.0894
R² = 0.86
c)
All:
y = 0.8981 + 0.6277
R² = 0.87
Spring:
y = 0.9293x + 0.3761
R² = 0.93
Autumn:
y = 0.8538x + 0.9244
R² = 0.79
d)
All:
y = 0.9127x + 6.5619
R² = 0.94
Spring:
y = 1.0618x + 6.7109
R² = 0.97
Autumn:
y = 0.9836x + 1.4589
R² = 0.90
e)
All:
y = 0.8646x + 1.0198
R² = 0.93
Spring:
y = 1.0945x + 0.7493
R² = 0.95
Autumn:
y = 1.0692x + 0.4448
R² = 0.90
f)
All:
y = 0.8642x + 0.267
R² = 0.83
Spring:
y = 0.9362x + 0.1322
R² = 0.80
Autumn:
y = 1.0409x + 0.105
R² = 0.89
g)
Spring
y = 1.0945x - 0.7493
R² = 0.9522
Autumn
y = 1.0692x - 0.4448
R² = 0.9039
All:
y = 0.9697x + 0.2746
R² = 0.79
Spring:
y = 1.0945x + 0.7493
R² = 0.95
Autumn:
y = 0.9955x + 7.0111
R² = 0.63