Na primeira etapa, é realizado a coleta de dados. A fim de serem utilizados dados de diversas fontes, é realizado uma busca por dados de caracteristicas socioeconômicas e ambientais, buscando assim a pluralidade da informação levantada.

Na segunda etapa, as principais *features* de interesse são escolhidas pelo usuário. Essa etapa pode contar com a ajuda de um usuário especialista, o qual saberá definir quais são as informações mais relevantes a serem consideradas dado o problema. Caso haja dados privados, esses podem ser também utilizados pelo usuário.

Na terceira etapa, ocorre o tratamento dos dados. Os dados categóricos são convertidos para dados numéricos e os dados numéricos são adequados conforme o algoritmo de clusterização a ser utilizado. Dessa etapa, é obtido como resultado a matriz de dados a ser utilizado com a fonte de dados para o algoritmo.

Na quarta etapa, é utilizado uma técnica de análise de componentes multivariadas. Essa etapa é importante para a metodologia adotada pois possiblita que a matriz multidimensional de dados seja convertida para uma matriz bidimensional, possibilitando assim a plotagem dos dados em um gráfico de dispersão bidimensional, por exemplo. Para o exemplo. Escolheu-se o *Principal Component Analysis* (PCA).

Na quinta etapa é realizada a escolha e aplicação de um algoritmo de clusterização. Para o exemplo, foi escolhido o algoritmo *Kmeans* por ser um algoritmo clássico e bem visto na literatura. Como escolheu-se essa técnica, foi necessário a adoção