1. The first line, “We’ve seen in the previous section the data we’ve collected and integrated in order to generate synthetic ecosystems.” can be changed to “In the previous section, we have seen the data being collected and integrated in order to generate synthetic ecosystems.” This way the “seen” verb is close to the object “data”.
2. Line 5 in page 6, “From this, we developed an R Package, ”. It is not clear what ‘this’ stands for. So may be replace it with something like “With this idea in mind, we developed an R Package,”.
3. Line 6 same page, “ Our goal for spew is that if the user could provide integrated data from the three required sources, spew would output a synthetic ecosystem.” What are the three required sources? I am guessing it is population counts, geographies and PUMS. But that is not very clear as you haven’t mentioned them as required sources before.
4. In the first paragraph of 4.1, the first sentence gives a high level idea of what spew is doing. But the second sentence on seems to be speaking about splitting a location into mutually exclusive regions. The next paragraph also talks about the same thing. So it might be a good idea to move the second sentence on to the next paragraph or incorporate them in the same paragraph. Also in the sentence, “ From this, we generate a synthetic population for each one of these regions.” the pronoun ‘this’ refers to what is not clear. You can even think of deleting the sentence as it is inherent in the next few lines. You could do something like, “At a high level, spew performs the function of taking our three integrated data sources, and outputs a synthetic ecosystem, see ?? for a demonstration. More specifically, spew works by splitting a location into mutually exclusive regions, the union of which adds up to the entire location. The PUMS data of any country usually includes a variable corresponding to a specific location, which is a superset of many smaller regions. We refer to this variable as the puma_id. Thus, each region in the PUMS data is typically subsetted to contain only data from the corresponding puma_id. (You can also replace this line with “Thus, the PUMS data is typically subsetted according to puma_id.” if it means the same thing.) Then we generate a synthetic population for each one of these regions. This leads to synthetic ecosystems which are more representative of the marginal distributions of each tract (location? region?).”
5. In “For example, in the United States we generate a unique synthetic population for each tract. In this case, we can think of each tract as one of our mutually exclusive regions. Note that each tract is contained within a Public Use Microdata Area (PUMA), and the United States PUMS data has a variable indicating which PUMA each record is located within. Thus, for each tract we subset the PUMS data to contain all samples from the particular PUMA the

tract is located in.”, note that in the first sentence readers might get confused with the usage of tracts which is a new term. It might be a better idea to start off with tract being one of the mutually exclusive regions and PUMA being the equivalent of the bigger location. So you can start with, “For example, in the United States a Public Use Microdata Area (PUMA) can be split into mutually exclusive regions called tracts.” This clears up the air in the beginning about what tract is and what PUMA is. Then you can go on to say “The PUMS data has a variable indicating which PUMA each record is located within. Hence for each tract, we can subset the PUMS data to contain all samples from the particular PUMA the tract is located in. From this subsetted data for each tract we can generate a unique synthetic population for each tract.” This way you follow the same structure as the previous paragraph which talks about the general scenario.

6. The next line onwards, “Once we have the correct PUMS data, …” talks about the steps once you have subsetted the data. This can go into a new paragraph as the previous paragraph and the current one mainly speaks about how to subset the data.
7. Page7, last line before the algorithm, “Thus, the default United States synthetic population has a subdirectory for each state, each PUMA within the state, and a synthetic population.” You should add “Thus, the default United States synthetic population has a subdirectory for each state, each PUMA within the state, and a synthetic population for every tract within the PUMA.”
8. For the algorithm it might be a good idea to put the caption above the algorithm. Otherwise it seems like the paragraph below the algorithm is being referred to as the algorithm.
9. In the paragraph below the algorithm, the third line “Also note that while the three required data-sources needed to generate the synthetic households and people, there is in principle no type of data, be it schools, workplaces, hospitals, mosquitoes, etc.., that we could not include into this framework.” is very confusing. It could be simplified and made clearer by changing the topic position. “Also note that while population counts, geographies and PUMS data are required, other types of data like data on schools, workplaces, hospitals, mosquitoes, etc.., could also be incorporated into this framework.”

The section is well written as decently structured. Some times you have used microdata and PUMS data interchangeably. It might be a good idea to stick to just one of the phrases. In some of the paragraphs (for example, second paragraph after algorithm on page 7) you have changed the tense you are writing in. It might be a good idea to stick to either present tense or past tense. That is just a suggestion. In general it sounds good even the way it is right now.