**Bacteria, Sand, and Water**

**Red**

**Summary**

A two-phase experiment was conducted to discover whether environments such as ground water, river water and piezometer water (a mix of ground water and river water) would affect colonization and growth of bacteria. This report presents an analysis using the experiment data. In general, there are differences among natural environments. River water has a significant lift of 4.54 pico moles per gram of sand in biomass and 441.28 pico moles per gram of sand per day in respiration comparing to piezometer water at initial, and the lift keeps significant throughout the experiment. Time trends in responses are negative in general although it is not supported with sufficient evidence. Model results provide the interaction between general water treatment and environmental condition does not have significant effect on the response trend. Among all factors of interests (environment, water treatment, time, interaction), environment has the largest effects on response variables, while the other factors also have effect on partial aspects. In the discussion session, a suggestion is made for future study.

**Background & Experiment Design**

Bacteria from different environments may have different colonization and growth. Environment and water treatment are factors of interest that affect bacterial behavior potentially. Researchers designed a two-phase experiment to test the influence. In the first phase, researchers put sterile sand into a sand infuser, putting it into one of three types of environments for a sex week incubation period. Such method was used to help bacteria colonize the sand. In the second phase, researchers used a glass receptacle (reactor) to contain each sample of sand. They applied one of four water treatments to the receptacle by adding water to the sand in the reactor. Researchers drained and refill the water each week, while measuring small subsamples of sand from each reactor in two different aspects weekly. First measurement happened at the beginning of second experiment phase.

**Variables:**

There are 378 observations, 7 variables in total from the data frame:

*Biomass* Response variable. It is the total mass of living matter in a given unit area. Final measurements are expressed in pico moles per gram of sand.

*Respiration* Response variable. It is a measure of the collective rate of metabolism of bacteria. Final measurements are expressed in pico moles per gram of the sand per day.

*Environment* It is the environment where the bacterial incubation happened. There are three environment types: ground water (GWS), piezometer water (PZS) and river water (RWS).

*Water* It is the water treatment that researchers applied to reactors. There are four water types: ground water (GW), river water (RW), artificial ground water (AGW) and artificial river water (ARW)

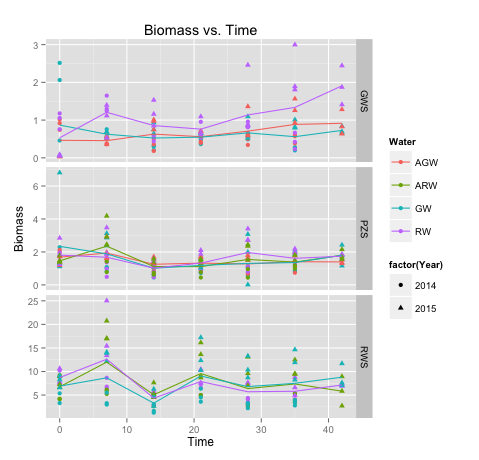
*Year* The experiment was first conducted in fall 2014. Then it was repeated in summer 2015. There are two levels in this factor variable: “2014” and “2015”.

*Time*  Number of week(s) until the response was measured. 2014 data have 5 weekly

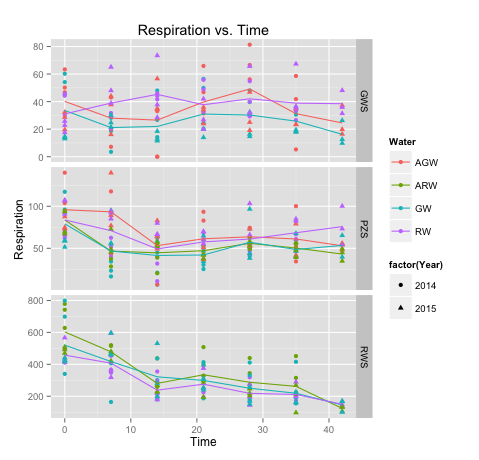
measurements. 2015 data have 6 weekly measurements.

**Exploratory Analysis**

Plots in this section helps find intuitions for whether there are (1) a general trend for biomass and respiration for all data (2) any difference across environment type or water treatment type (3) other unanticipated difference. Biomass values and respiration values from three environment types are presented in three layers in the plot respectively. Colors are used to mark different water types. Scales are adjusted based on values for a clearer trend observation. Lines connect the mean value for each water type within the given environment type. Observations from different years are separated by shape.



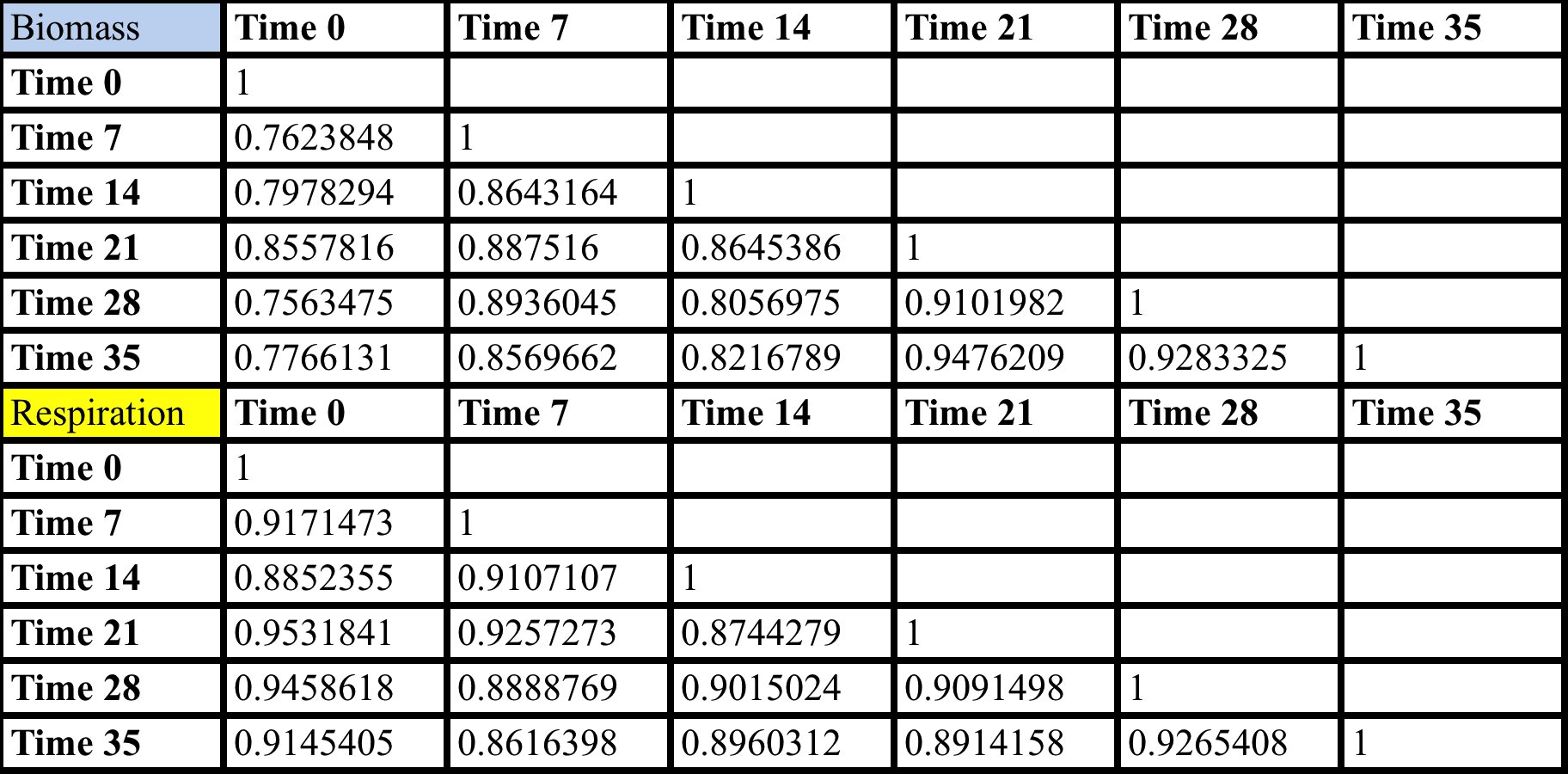
In the plot for biomass, RWS has higher mean value and variation in biomass, as the scale in RWS is larger than other two layers. Higher initial points may attribute to that river water is friendlier for bacteria growth or there is high bacteria density in river water. But the fact that higher biomass in following time points supports the first hypothesis. RW line in GWS layers has an obvious lift from other lines. This could imply the interaction between RW and GWS leads to a higher biomass compare to interactions between other water treatments and GW. Also, the interaction between Time, GWS and RW may have a significant lift in biomass trend. Lines in PZS layers do not show essential difference in either trend or biomass value.



For respiration, there is a clear downward trend in RWS layer. This could imply that bacteria in river water have decreasing respiration as time increases, while respiration from other environment is barely affected by time because trends in GWS and PZS look constant. Same as biomass, respiration is much larger in RWS than in GWS and PZS. There is no significant difference of lines within same layer, which indicates water treatment may have a weak effect on respiration.

In both biomass and respiration plots, we can find that most triangular points are higher than dot points. Triangles are response observations from year 2015. Responses from 2015 data set are clearly larger than responses from 2014. Therefore, I decide to treat “Year” as a fixed effect in modeling process in order to point out the difference caused by different years.

Table below indicates correlation among responses:

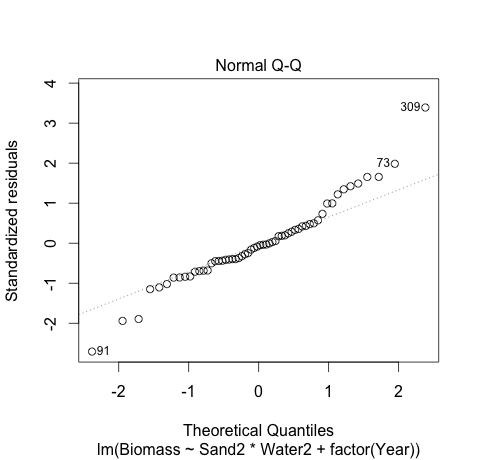
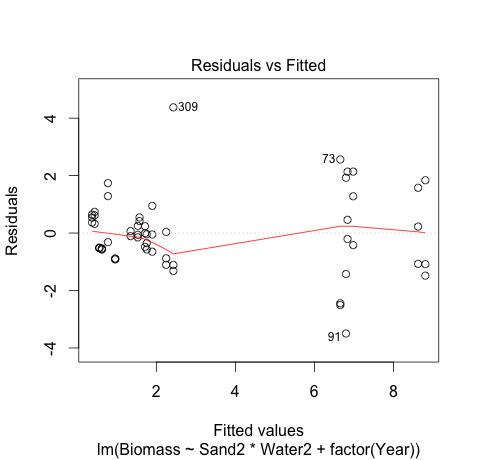


For both responses, there is strong correlation among each observation, which means each response value is highly influenced by others. The correlation is not affected obviously by time intervals. During the modeling process, we should pay attention to correlation influence.

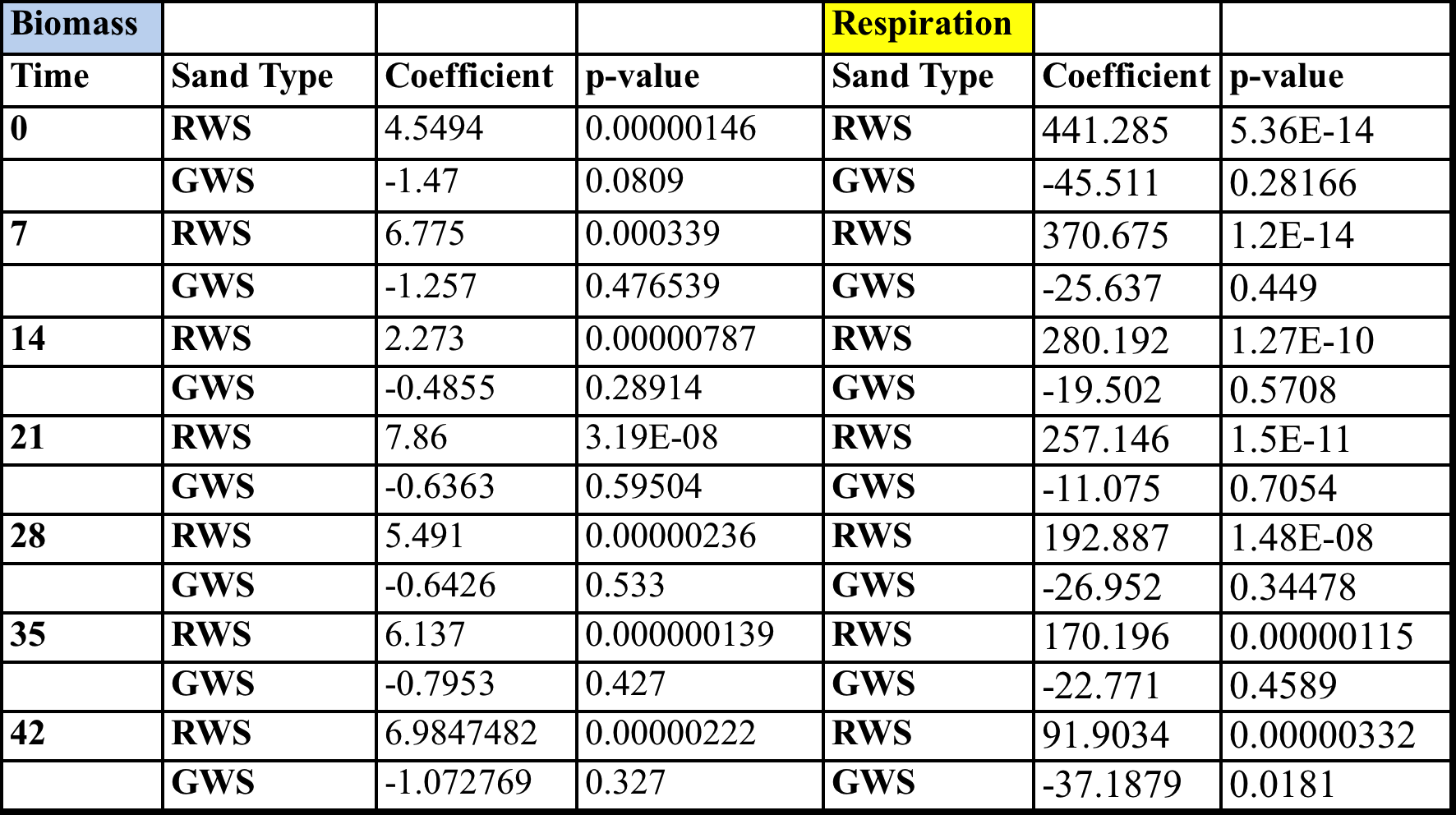
**Modeling Process**

Because response observations are not independent, we can either separate data by time point then use ANOVA to test the difference, or use a mixed effect model to fit data. For the first method, comparison within in each time point could eliminate the influence caused by time correlation. In that case, we will have seven comparisons for each response, including the initial measurement, with ANOVA test. The result is presented in Appendix. From those ANOVA tables, environment shows significant difference in all ANOVA tables, which indicates that we have evidence to support there is variation caused by environment type at every time point.

Next, we can use linear model to calculate the exact difference amount. We would like to test the effect from environment, water treatment type and their interaction. Moreover, the effect from year difference cannot be ignored according the exploratory analysis. The linear model is



The above residual plots and Q-Q plot indicates OLS assumptions are met in this model. Table below will specify the significant differences among the natural environments (SandRWS, SandGWS, SandPZS)：



Coefficient with p value smaller than 0.05 will be considered significant. Only RWS has significant positive variables, which indicates there is sufficient evidence to say that RWS biomass is larger than biomass of baseline (PZS). Using biomass data in the first row as an example to read the table: At time 0, the combination of RWS and GW has 4.55 pico mole per gram of sand biomass higher than the combination of PZS and GW in the same unit(combination of PZS and GW is the baseline in the model). Due to the insignificancy of other water type, we can conclude that among sand types, river water has more biomass than piezometer water. Although difference between GWS and PZS is not significant in this test, we find GWS has mean values all lower than PZS. From the exploratory study, we can also see lines of GWS biomass cluster in a lower level than those of PZG. The insignificancy could be caused by large standard deviation, which leads to a small t score and large p value. The test result for respiration is similar to biomass. Bacteria from river water have higher respiration than Bacteria from piezometer water at every time point. No obvious evidence to indicate there is difference between piezometer water and ground water although all the respiration values for GWS is lower than PZS.

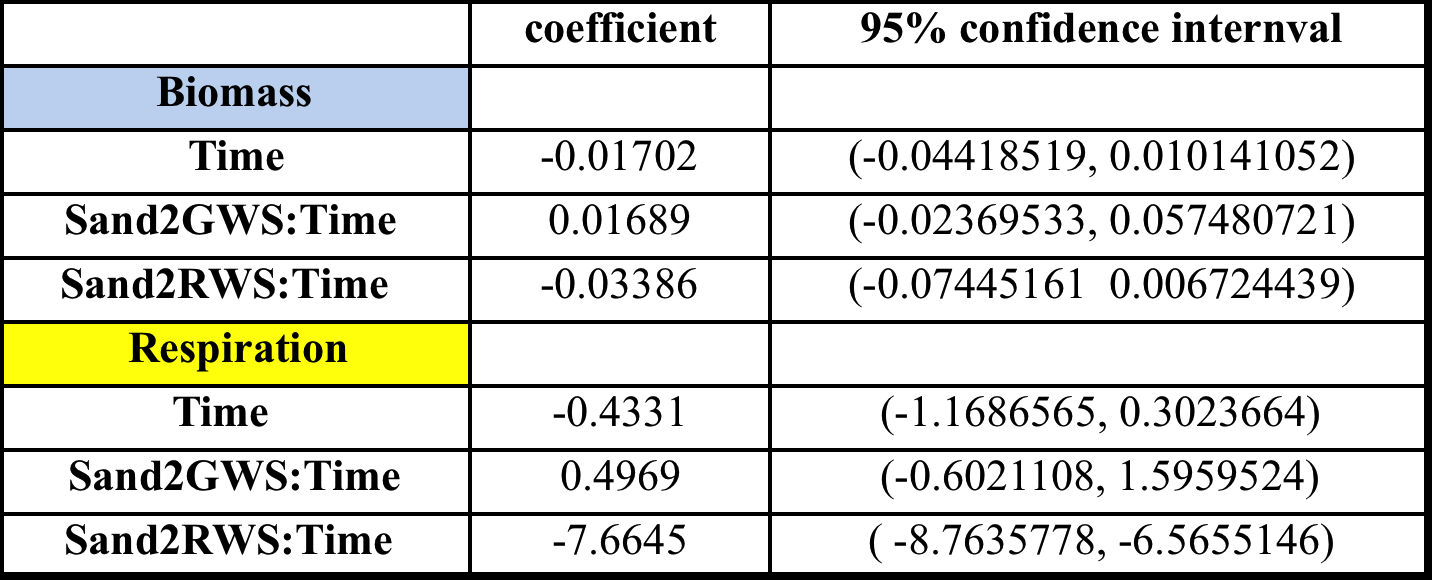
We use the following mixed-effect model to testify time trends of response variables:

To test whether all variables and interactions are significant in the full model, we can use AIC backward elimination for variable selection. The final models are:

(Model 1)

(Model 2)

Coefficients related to response time trend are listed in the following table. Full model summary tables are in Appendix.



Biomass has a downward trend in general. Ground water environment may declines less in biomass trend and river water environment accelerates the decline speed. We need more sufficient evidence to support the conclusion above but at least we have some intuition about effect on time trend. For respiration data, the table shows bacteria from river water have a slope 7.66 significantly lower than other bacteria type. This means bacteria from river water will decrease faster in respiration. This conclusion can also be detected from the ANOVA tests above. RWS has significantly higher respiration, but the coefficient is decreasing, which implies a downward trending of RWS biomass because the baseline biomass is constant. Both models do not conclude water type, which means time trend will not be influenced by water treatment, neither the interaction between water treatment and environment.

**Discussion:**

To answer client’s questions:

For the first question, there are difference among the natural environment with regard to the biomass and activity of bacteria that colonized the sand. The differences are quantified by the difference among biomass and respiration at each measurement time point. River water has larger biomass and respiration than responses from other two environment types at the initial point. This is the result from ANOVA comparison.

For the second question, time trends are generally negative in both biomass and respiration. From insufficient evidence, we find ground may decrease the speed of decline in biomass while river water increases it. Result also indicates with sufficient evidence that as time increase, river water bacteria will decrease faster in respiration than other two groups. Evidence is not enough to support there is more growth when the water treatment matches the environmental condition because water treatment is not proved to be effective on time trend. An unexpected discovery is that interaction between RWS and RW will have a decrease in biomass change.

For the third question, after having a consulting session with the clients, I find it is difficult to identify which factor has the “largest effects” because we do not have a standard scale to measure the effect. Therefore, I will make the decision based on the comprehensiveness of factor influence. Water treatment has no essential effect on both values and time trend for both responses. Time is influential in time trend of response variables, but effect from most of environment and water type combination does not have a clear upward or downward trend. Therefore, environment is the factor that has “largest effect”. The difference among the natural environments with regard to the biomass and respiration is obvious, and it also contributes to time trend of response variable. Environment has the most comprehensive effect in this experiment.

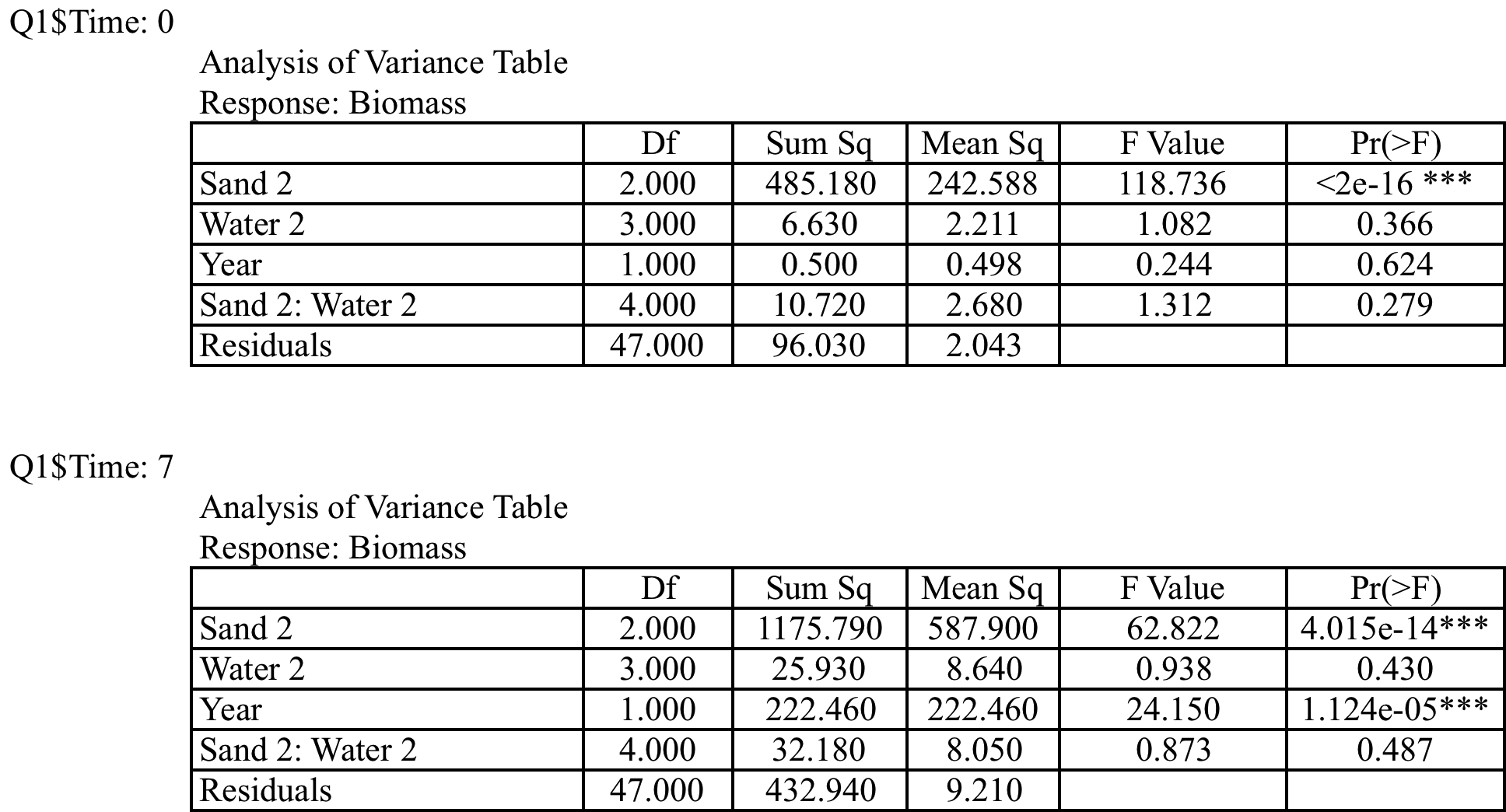
For future improvement, I suggest experiment designers to conduct repeat experiment in the same season as the original experiment. Plots shows that data in 2014 and 2015 are difference in exploratory analysis, so I treat “Year” as a fixed effect factor in following models. In fact, “Year” does have significant influence on responses. Data from 2015 have significant higher values in biomass and respiration than data from 2014. One reasonable explanation could be that the repeated experiment is conduction in difference season (the first experiment is in fall and the second is in summer). Difference in season may cause difference biological situation for bacteria.

Another suggestion is increasing sample size as large as possible. In this experiment, we could have more conclusions about difference across environment and water time if we have less variation in the data. Increasing sample size is the most direct way to solve this problem and provides us more evidence. Also, recording biomass and respiration in a shorter time interval may help to detect a clearer time trend in responses.

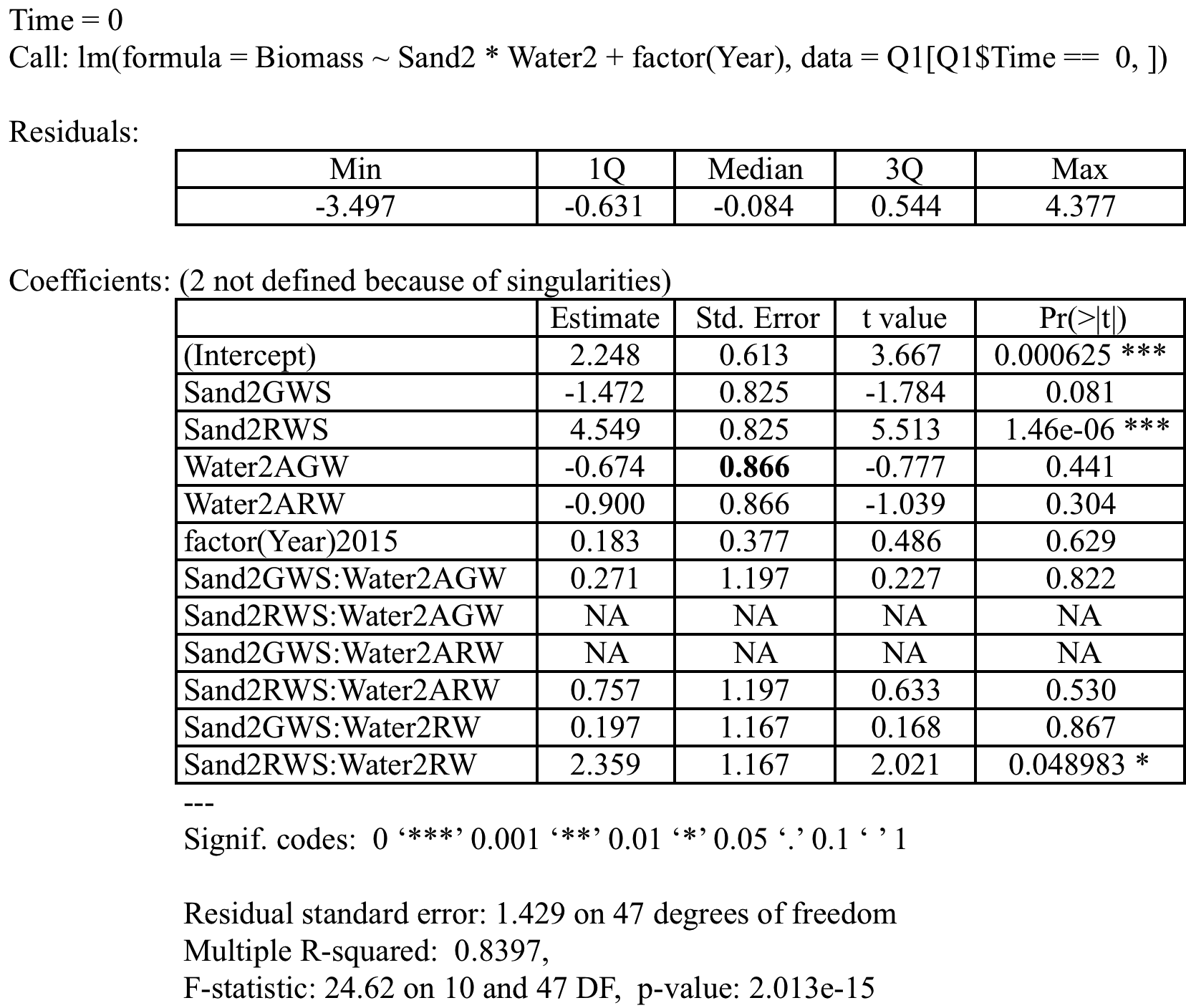
Finally, I have a doubt for the method of collection samples at the first stage. In this experiment, biomass and respiration in river water is prominently higher than the other group. If the researcher put sand infuser into three types of water in the nature for a six-week incubation, then it will be more bacteria collected from the rush river water than still ground water. The substantially difference in the initial point could mislead the experiment result potentially. In the future study, research could improve the sample collection method, trying to balance the initial value. It will help to figure out a more precise result.

**Appendix**

**1. Partial ANOVA test in Q1(biomass)**

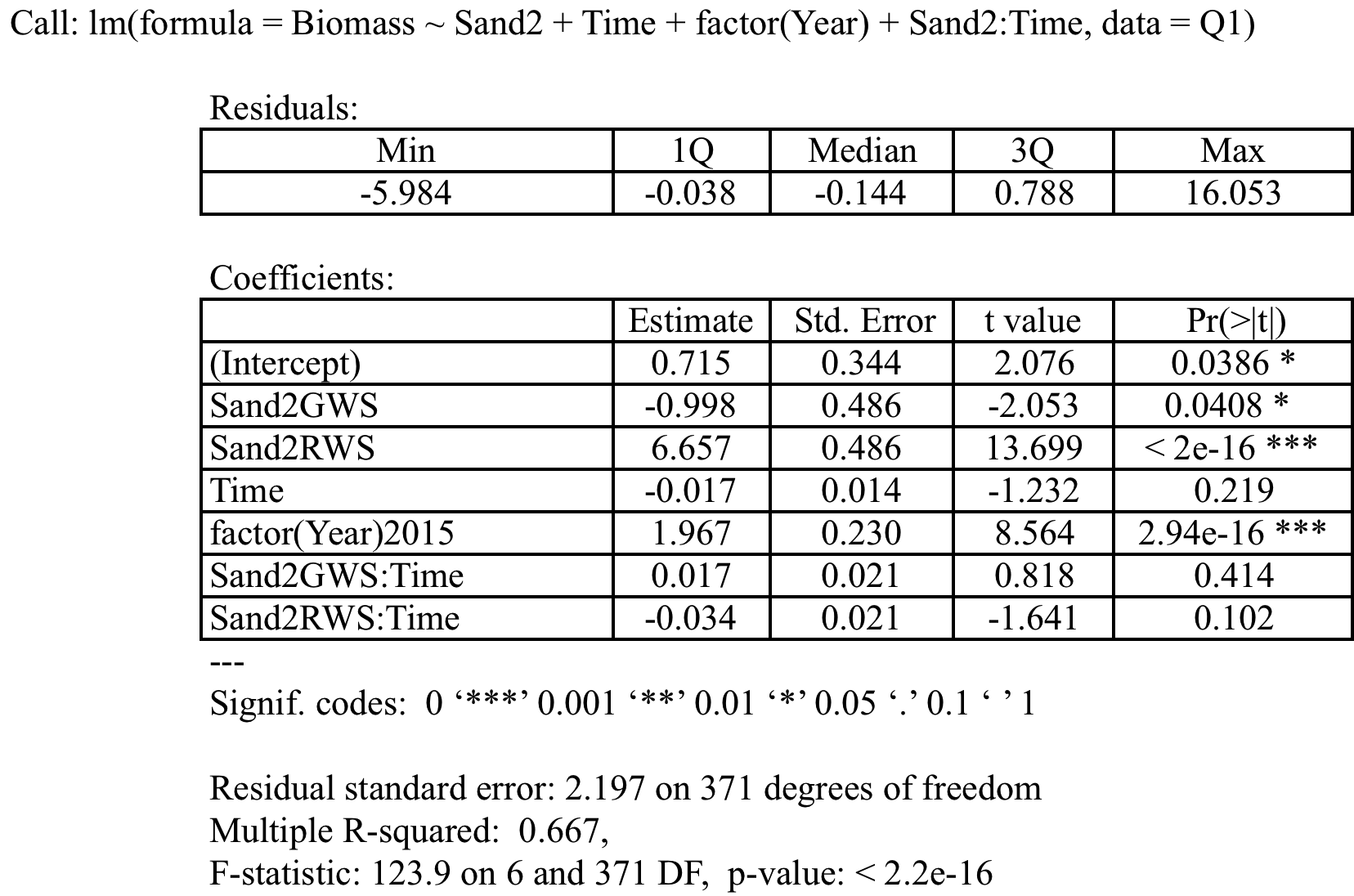


**2. Partial Summary Table in Q1:**



**3. Summary table for model in Q2**

**Biomass**



**Respiration**

